인간 질병에서 DNA 메틸화 지역의 고차상호작용 탐색을 위한 진화적 연관관계 학습

이제근 O1 김수진 1 장병탁 1,2
1 서울대학교 생물정보학 협동과정
2 서울대학교 컴퓨터공학부
{jkrhee, sjkim, btzhang}@bi.snu.ac.kr

Evolutionary association learning for detecting higher-order interactions of DNA methylation regions in human diseases

Je-Keun Rhee O1, Soo-Jin Kim1, Byoung-Tak Zhang1,2
1Interdisciplinary Program in Bioinformatics, Seoul National University
2Department of Computer Science & Engineering, Seoul National University

요 약

DNA 메틸화는 후성유전학의 한 유형으로 유전자 발현을 조절하여 질병을 비롯한 다양한 생물학적 프로세스에 영향을 준다고 알려져 있다. 따라서 DNA 메틸화 정도와 인간 질병과의 연관성에 관한 연구는 질병의 원인 및 기전을 밝히고 메틸화 프로세스 조절을 통한 질병 치료 방법 개발을 위한 기반이 될 수 있다. 유전자 발현 조절 및 질병 발생은 많은 인자들의 복합적인 상호작용에 영향을 받으므로, 여러 위치에서의 메틸화 정도들의 고차원 조합을 이용한 질병과의 연관 관계 분석이 필수적이다. 본 연구에서는 진화 연산과 가중치 학습에 기반하여 유방암 발생과 연관되어 있는 메틸화 위치의 고차 상호작용을 탐색할 수 있는 방법을 제안한다.

1. 서론

최근 다양한 형태의 생물학 데이터가 대량으로 생산되고 있으며, 이를 이용하여 전역 유전체 수준에서 유전적 변이를 분석하고 질병의 유전적 원인이 되는 인자들을 찾거나 질병 발생 가능성을 예측하려는 시도가 다수 이루어지고 있다. 특히 후성유전학(epigenomics)은 DNA 서열 자체에는 변화가 없지만 유전자 발현이나 질병 등의 표현형에 영향을 미칠 수 있는 기작에 대한 것으로, 이 중 가장 대표적인 것이 DNA 메틸화(DNA methylation) 관련 연구이다. DNA 메틸화는 유전체 서열에서 시토신(cytosine)에 메틸기가 결합되는 것으로, DNA 메틸화가 많이 되면 유전자 발현을 억제할 수 있다고 알려져 있다[1].

과거의 DNA 메틸화 관련 연구에서는 특정 한 영역의 메틸화 정도와 이와 관련된 한 유전자 발현양 변화를 실험적으로 관찰하여 DNA 메틸화 기작의 생체 내 영향을 연구해왔다. 하지만 최근 전역 유전체 수준에서 DNA 메틸화 정도를 측정할 수 있는 기법들이 발달됨에 따라, 보다 거시적인 관점에서 DNA 메틸화 정도와 이에 따른 세포 내 기작 변화와 질병 발생의 상관 관계 등을 관찰하는 것이 가능해지게 되었다[2]. 특히 질병을 포함한 대부분의 생체 내 기작들은 각 인자들이 독립적으로 영향을 미치기 보다는 수 많은 인자들의 복합적 상호작용에 의해서 유발되므로, 복잡한 질병 발생 원인을 보다 명확히 밝히기 위해서는 여러 인자들을 조합적으로 고려한 고차원적인 분석이 필수적이다. 하지만 전역 유전체에서 DNA 메틸화가 일어날 수 있는 위치는 매우 많으므로 이들 인자들의 고차원 조합에 대한 모든 경우의 수에 대한 조합적 인자 공간(combinatorial feature space)을 완전히 탐색하여 분석하는 것은 현실적으로 불가능하다.

따라서 전역 유전체 수준에서 각 인자들 간의 관계를 분석하고, 인자들의 다중 조합과 질병과의 관련성을 분석할 수 있는 새로운 방법 개발이 필요하다. 본 논문에서는 기계학습 기술에 기반한 특정 질병 관련 다중 메틸화 인자 조합 분석을 위한 진화적 연관관계 학습 방법을 제안한다. 이에 본 논문에서는 유방암 관련 대용량 DNA 메틸화 데이터를 이용하여 제안한 방법의 분류 성능을 확인하고, 유방암에 연관되어 상호작용하는 것으로 생각할 수 있는 DNA 메틸화 영역의 고차원 조합을 찾는다. 이를 통해 제안한 방법이 기존의 다른 분류 모델과는 달리 인자들의 고차원 상호작용을 탐색하여 분류에 중요한 인자 조합을 찾을 수 있다는 것을 결과로 제시한다.

2. 진화적 연관관계 학습

2-1. 실험 방법

그림 1은 본 논문에서 제안하는 방법의 흐름도를 나타내며, 전체적인 알고리즘은 다음과 같이 구성된다.

a) 주어진 염색체(chromosome)의 길이(order) l에 따라 인자 수를 랜덤으로 조합하여 모집단 개체 크기(population size) s 개의 개체(individual)들을 생성한다.

b) 모든 개체들에 대해 랜덤으로 가중치(weight) w_j $(0 \le j \le s)$ 값 초기화한다.

c) 전체 n 개로 구성된 학습 데이터 $\mathbf{x}$에서 i $(0 \le i \le n)$ 번째 인스턴스(instance) $\mathbf{x}_i$에 대해 각 염색체의 가중치 값을 이용하여 결과값 $f(\mathbf{x}_i)$을 예측한다. 여기서 x_{ijk}은 i 번째 학습 데이터에서 개체 j에 속하는 인자 중 k $(1 \le k \le l)$ 번 째 값을 의미한다.

$$f(\mathbf{x}_i) = \begin{cases} 1 & if \ \sum_{j=0}^{s} w_j \cdot x_{ij} > 0 \quad (x_{ij} = \frac{x_{ij1} + \cdots + x_{ijk} + \cdots + x_{ijl}}{l}) \\ -1 & else \end{cases}$$

d) 예측된 결과값 $f(\mathbf{x}_i)$와 현재 인스턴스 $\mathbf{x}_i$ 의 실제 클래스 값 t_i의 차이를 이용하여 다음 규칙을 통해 각각의 가중치를 차례로 업데이트한다. 여기서 η는 학습 속도(learning rate)이다. 본 모델에서의 w_j은 일반적인 진화 연산에서 사용하는 적응도(fitness) 값으로 해석될 수 있다.

$$w_j = w_j + \eta(t_i - f(\mathbf{x}_i))x_{ij}$$

e) 수렴할 때까지 에포크(epoch) 수 c 만큼 c)-e)번 과정을 반복하면서 학습 과정이 수렴되도록 한다.

f) 다음 식을 이용하여 모든 개체들에 대하여 가중치 w_j의 절대값을 기준값 θ $(\theta \ge 0)$ 와 비교하여 모든 개체들에 대해 $e(w_j)$의 값을 결정한다.

$$e(w_j) = \begin{cases} -1 & if \ |w_j| < \theta \quad (\theta \ge 0) \\ 1 & else \end{cases}$$

g) $e(w_j)$의 값이 1인 개체들은 그대로 다음 세대(generation)으로 넘긴다 (elitism). 한편, $e(w_j)$의 값이 -1인 개체들의 수만큼 랭킹 기반 선택 방법(ranking selection)에 의한 교차(crossover) 연산과 랜덤 생성(random generation)을 통해 다음 세대의 개체들을 각각의 정해진 비율 λ $(0 \le \lambda \le 1)$에 따라 재생성(reproduction)한다. $e(w_j)$의 값이 -1인 개체 수를 γ이라고 할 때, 교차연산에 의해 생성되어야 할 개체 수 α와 랜덤으로 생성되어야 할 개체 수 β는 다음과 같이 계산된다.

$$\alpha = \lambda(s - \gamma)$$
$$\beta = s - r - \lambda(s - \gamma)$$

돌연변이(mutation) 연산도 매 세대마다 사전에 정의된 비율 m $(0 \le m \le 1)$에 따라 일어난다.

h) 사전에 정의된 최대 세대 수 g에 따라 종료 조건을 만족할 때까지 c)-h)의 과정을 반복하면서 세대에 진행에 따라 개체들을 진화시킨다.

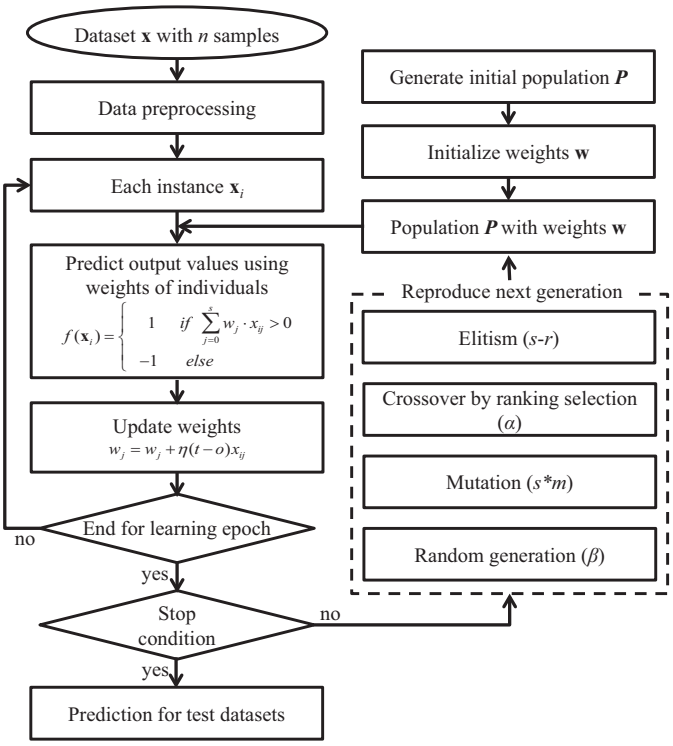

그림 1. 제안된 알고리즘에 대한 흐름도

2-2. 실험 데이터

본 논문에서는 영국 UCL (University College London)의 Joanna Zhuang의 연구진에 의해 만들어진 DNA 메틸화 데이터를 이용하여 실험하였다[3]. 이 데이터는 유방암 및 정상 조직에서 Illumina Infinium 27k Human DNA methylation Beadchip을 이용하여 27,000여 개의 CpG 사이트에서의 DNA 메틸화 정도를 측정한 것이다. 본 논문에서는 전체 유전체 중에서 17번 염색체에 존재하고 있는 1,586개의 메틸화 위치 데이터를 이용하여 전체 137개의 샘플 중 일부 값이 누락된 샘플은 제거하고 실험을 수행하였다.

3. 실험 결과

본 논문에서 보여주는 결과는 표 1의 설정에 따라 실험하여 얻은 결과이다. 표 2는 제안한 방법과 다른 방법의 분류 성능을 제시한다. 모든 결과는 표 1에 나온 동일한 파라미터 설정 하에서 l의 값만 변화시켜보면서 10-fold cross-validation을 통해 얻은 결과이다. 본 논문에서 제안된 방법은 다른 방법과 유사하거나 보다 높은 분류 정확도를 보인다는 것을 알 수 있다.

표 1. 파라미터 설정

파라미터	값	파라미터	값
l	2-10	λ	0.2
s	10,000	m	0.001
θ	1.0	c	50
η	0.1	g	4

표 2. 분류 성능

알고리즘	분류 정확도(%)
제안된 모델(*l*=2)	94.95
제안된 모델(*l*=3)	95.96
제안된 모델(*l*=4)	94.95
제안된 모델(*l*=5)	94.95
제안된 방법(*l*=10)	92.93
SVM (Poly. kernel)	96.97
RBF network	91.92
Naïve Bayes	93.94
Decision Tree	88.89

표 3 유방암 관련 상호작용 메틸화 위치 (*l*=4)

순위	메틸화 위치	관련 유전자	가중치
1	cg17233506	HOXB1	7.664
	cg24164563	FOXJ1	
	cg14315198	UNC119	
	cg15799267	ALOX15B	
2	cg15238200	TRIM65	7.184
	cg15786837	HOXB13	
	cg12989650	ARHGEF15	
	cg09040752	AOC3	
3	cg21846903	VTN	-6.875
	cg16585619	KRT19	
	cg04947157	TMC6	
	cg24392479	ASPSCR1	
4	cg25145670	HOXB4	6.796
	cg19895197	C17orf37	
	cg21546671	HOXB4	
	cg08089301	HOXB4	
5	cg20723355	FBXO39	6.768
	cg13053608	LGP1	
	cg04216597	CACNA1G	
	cg04106785	CDK5R1	

표 3은 *l*의 값이 4일 때 반복 실험을 통해 얻어진 가중치가 높은 상위 다섯 개의 상호작용하는 메틸화 위치이다. 가장 순위가 높은 상호작용 위치들은 HOXB1, FOXJ1, UNC119, ALOX15B 등의 유전자와 관련된 위치이다. 즉 DNA 메틸화가 유전자 발현에 직접적으로 영향을 미칠 수 있으므로, 이 영역의 메틸화 정도에 의해 해당 유전자의 발현양이 달라질 수 있다는 것을 의미하며, 이 영역들의 메틸화 정도가 동시에 변화하는 것이 유방암 발생과 관련될 수 있다는 것을 의미한다. 이 중 ALOX15B (15-LOL-2)의 경우 유방암 환자에서 발현양이 낮아진다는 사실이 이미 알려져있다[4]. UNC119의 경우 그 기능이 아직 확실히 알려져 있지는 않지만 이 유전자와 물리적으로 상호작용할 수 있는 ARL2 유전자가 유방암 세포에서 영향을 준다는 사실은 알려져 있으므로, UNC119 유전자 역시 직간접적으로 유방암 세포에서 기능을 할 가능성이 있다[5]. 또한 HOXB1 유전자는 Hox 계열 유전자들 중 하나로 이들 유전자들이 암 발생과 밀접하게 관련되어 있다는 사실은 이미 보고된 사실이며, 이에 최근 이들 유전자들의 후성유전학적 발현 조절과 암 발생과의 관계에 대한 연구도 활발히 진행되고 있다[6]. 특히 FOXJ1의 경우, DNA 메틸화 기작 변화 실험을 통해 후성유전학적으로 발현양을 조절하는 것에 의해 이 유전자가 유방암 억제 유전자로 기능을 할 수 있다는 연구가 발표되기도 하였다[7]. 또한 HOXB4 유전자의 서로 다른 위치에서의 메틸화가 함께 상호작용한다는 것이 본 논문에서 제안된 방법에 의해 상위에서 발견된 실험 결과를 통해 본 방법이 서로 관련한 상호작용할 가능성이 높은 메틸화 지역들 탐색에 유용하다는 것을 보여준다. 이와 동시에 본 결과는 유방암 관련 연구에서 HOXB4 유전자의 기능과 메틸화에 의한 발현 조절에 대해 보다 면밀히 연구해볼 필요성도 제시한다.

4. 결론

본 논문에서 유방암과 관련된 DNA 메틸화 지역의 고차원 상호작용을 찾기 위한 진화적 연관 관계 학습 방법을 제안하였다. 본 방법은 암과 정상군을 높은 성능으로 분류하면서, 동시에 생물학적으로도 의미있는 DNA 메틸화 지역의 고차상호작용을 찾을 수 있었다. 향후 보다 다양한 데이터를 이용한 추가 실험을 통해 보다 생물학적으로 유의한 의미있는 질병 관련 DNA 메틸화 위치 조합을 찾을 수 있을 것으로 기대한다.

감사의 글

본 논문은 교육과학기술부 국가연구재단의 지원을 받아 수행된 연구(No. 2012-0005643)에 의해 지원되었음.

참고문헌

[1] Suzuki, M. and Bird, A., DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476, 2008.

[2] Laird, P. W., Principles and challenges of genome-wide DNA methylation analysis, *Nat. Rev. Genet.* 11, 191-203, 2010.

[3] Zhuang, J., *et al.*, The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer., *PLoS Genet.*, 8(2), e1002517, 2012.

[4] Jiang, W. G., *et al.*, Reduction of isoforms of 15-lipoxygenase (15-LOX)-1 and 15-LOX-2 in human breast cancer. *Prostaglandins Leukot Essent Fatty Acids* 74, 235–245, 2006.

[5] Beghin, A., *et al.*, ADP ribosylation factor like 2 (Arl2) protein influences microtubule dynamics in breast cancer cells, *Exp. Cell Res.*, 313, 473–485, 2007.

[6] Shah, N. and Sukumar, S., The Hox genes and their roles in oncogenesis, *Nat. Rev. Cancer* 10, 361-371, 2010.

[7] Demircan B, *et al.*, Comparative epigenomics of human and mouse mammary tumors. *Gene Chromosome Canc.* 48, 83– 97, 2009.