유전 알고리즘을 이용한 천식과 단일염기다형성(SNP)의 연관성

임상섭*, 김승현**, 위규범*
*아주대학교 정보통신전문대학원
**아주대학교 의과대학
e-mail:{leemss, kimsh, kbwee}@ajou.ac.kr

Detection of SNPs involved in the development of asthma with genetic algorithms

Sangseob Leem* , Seung-hyun Kim** , Kyubum Wee*
*Graduate School of Information & Communication, Ajou University
**School of Medicine, Ajou University

요 약

천식(Asthma)과 같은 복합질환(Complex Disease)의 원인과 작용 모델을 찾기 위해서 여러가지 통계적인 방법들과 기계 학습(Machine Learning)의 방법 등이 사용되고 있다. 본 연구에서는 유전 알고리즘을 이용하여 천식 환자와 대조군들을 분류할 수 있는 단일염기다형성(SNP, Single Nucleotide Polymorphism)의 조합에 대하여 조사한다.

1. 서 론

천식(Asthma)과 같은 복합질환(Complex Disease)의 원인과 그 작용 모델을 찾기 위해서 여러 가지 통계적인 방법들과 기계학습(Machine Learning)의 방법들이 사용되고 있다. 통계적인 방법들은 수학적으로 원리와 타당성이 증명되어 있지만 분석해야 하는 자료의 수가 많아지면 통계적 모델을 찾는 시간이 늘어 거의 불가능하게 된다. 예를 들어, 로짓 회귀 분석(Logistic Regression)의 방법을 이용할 경우 입력 인자들의 수가 많아져서 다차원 분석을 해야 하므로 인자의 수가 늘어나면 거의 불가능 하다[1]. 또한, p-value, 엔트로피 등을 이용하여 모든 SNP의 조합을 계산해 보기 위해서는 O(n!)의 시간복잡도를 가지기 때문에 후보 SNP의 수가 증가하면 계산하기 어려운 단점이 있다. 인공신경망(Neural Network)을 이용한 방법은 SNP의 수가 늘어나도 그 계산시간이 크게 늘어나지 않고, 예측 정확도가 일반적으로 높은 장점이 있다[2,3,4]. 그러나 인공신경망을 이용해 찾은 판단 모델이 겉으로 드러나지 않는 가려진 모델(Black Box)이기 때문에 작용 기작(Mechanism)을 알아보기 힘들다. 결정트리(Decision Tree) 방법은 규칙을 알아보기 쉽고 계산시간도 비교적 빠른 편이지만, 데이터의 미세한 변화에도 판단 모델의 전체적인 모양이 크게 바뀌는 문제점이 있다[5,6]. MDR(multifactor dimensionality reduction) 방법은 적용모델을 결정하지 않고 사용할 수 있다는 장점이 있지만, 모든 조합을 시험해 보기 위해서 많은 시간이 소비되고 다차원 분석을 할 경우, 분할표(Contingency table)에서 비어 있는 원소가 발생할 확률이 높아지는 문제점이 있다[7].

본 연구에서는 복합질환인 천식 환자와 대조군의 자료를 이용하여 환자(case)와 대조군(control)을 구분하기 위하여 유전 알고리즘을 이용한 방법을 실제 SNP 자료들에 적용하여 천식 발병에 중요한 요인이라고 판단되는 SNP를 찾아보았다.

2. 본 론

J. Moore 등의 연구에서는 셀룰러 오토마타(CA)를 SNP의 조합과 작용 기작을 찾는 모델로 사용하고, 우수한 CA를 찾기 위하여 유전 알고리즘을 사용하였다[8].

이 논문은 2006년도 정부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2006-000-10775-0)

2.1 Cellular Automata(CA)

2.1.1 일반적인 CA

CA는 세포(cell)들이 각각 상태(state)를 가지고, 이웃하는 세포들의 상태와 변화 규칙(rule)에 따라서 다음 상태가 결정되는 모델이다. 간단하게는 1차원의 배열로 이루어질 수도 있고, 2차원 또는 그 이상의 차원으로 이루어 질 수 있다. 이 연구에서 CA는 SNP의 표현형(Genotype)을 입력으로 받아서 정해진 반복횟수(iteration) 동안 상태가 변화한 후 특정한 세포의 상태가 환자와 대조군을 구분하는 기준이 된다.

2.1.2 이전 연구에서의 CA

Moore의 연구에서 하나의 CA는 다섯 개의 세포와 1개의 반복횟수, 규정표(rule table)로 이루어진다. 세포들은 1차원으로 이루어지고 각 세포들은 네 가지의 상태 중 하나의 상태가 된다. 첫 번째 상태는 우성/우성(major homogeneous), 두 번째 상태는 우성/열성(heterogeneous), 세 번째 상태는 열성/열성(minor homogeneous)을 나타내고 네 번째 상태는 해당 SNP가 아무 의미가 없음을 나타내는 것으로 다섯 개로 고정되어 있는 SNP의 개수가 5개 이하인 경우까지 판단할 수 있도록 되어 있다.

2.1.3 본 연구에서 고안한 CA모델

이 연구에서 CA는 다음과 같이 구조화 되었다.

SNP1	SNP2	SNP3	SNP4	Iteration	Rules
Selector1	Selector2	Selector3	Selector4		

(그림1) CA 구조

1) 세포의 개수가 5개로 고정되어 있던 Moore 등의 방법에서 세포의 개수가 가변적으로 바뀔 수 있도록 바꾸고 여러 가지 개수를 테스트 해 보고 그 최적 개수를 찾아보았다.

2) SNP1 ~ SNP4는 해당 세포에 입력으로 들어올 SNP자료번호이다.

3) 각 SNP입력에 Selector를 두어 입력된 SNP의 표현형(genotype)이 작용하는 모델을 우성(dominant)과 열성(recessive) 중 하나의 모델을 따르도록 하였다. 우성 모델일 경우는 우성/우성, 우성/열성이 1로 표현되고 열성/열성이 0으로, 열성 모델일 경우는 우성/우성이 1, 우성/열성과 열성/열성이 0으로 표현되었다.

4) Iteration은 반복횟수이다. Moore 등의 연구에서는 128이하의 수로 제한되었지만, 실제 자료들을 가지고 테스트 해 보았을 때, 세포 수의 두 배로 제한을 낮추었을 때와 성능의 차이가 없고 CA를 구성하는 시간이 절약되기 때문에 세포 수의 두 배로 제한하였다.

5)Rules는 규정표를 가지고 있다. Moore 등의 연구는 64개의 규정을 가지고 있어야 하지만 본 연구에서는 Seletor를 도입하였기 때문에 8개의 규정만 가지고 있으면 된다.

2.1.4 CA의 동작

CA의 동작을 설명하기 위해서 실제 발생할 수 있는 CA의 예와 함께 설명하겠다. 일단 입력으로 들어오는 SNP의 개수가 4인 경우에 아래 그림과 같은 CA가 선택된 경우의 예를 보이겠다.

SNP1 2	SNP2 5	SNP3 8	SNP4 7	Iteration 5	Rules 01010101
Selector1 Dominant	Selector2 Dominant	Selector3 Recessive	Selector4 Recessive		

(그림2) SNP 개수가 4일 때 CA의 예

위의 예는 SNP입력으로 2,5,8,7이 입력으로 선택되었고, 2번/5번 SNP는 우성모델, 8번/7번 SNP는 열성모델로 선택되었고, 반복횟수는 5, 규정표는 01010101로 선택되었을 때의 예이다. CA의 동작을 예로 보이기 위해 한 사람의 자료가 2번 SNP는 우성/열성, 5번 SNP는 열성/열성, 8번 SNP는 우성/우성, 7번 SNP는 우성/열성이라고 가정하고 동작을 설명하겠다.

1) 한 사람의 해당하는 SNP 자료를 Selector의 형태에 맞게 가져온다. 예에서는 다음 그림과 같은 결과가 된다.

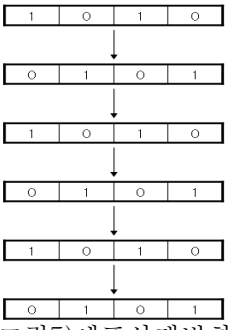
1	0	1	0

(그림3) SNP 입력 결과

2) 반복 횟수만큼 규정표의 규정들을 적용시킨다. 규정표가 01010101인 경우, 그 규정표를 알아보기 쉽게 다시 표현하면 (그림4)와 같아진다. 해당 세포를 포함하여 좌우 세 개의 세포 값에 따라서 해당 세포의 다음상태가 결정된다. 동작하는 과정을 살펴보면 5번 상태가 변화는 과정을 살펴보면 (그림5)처럼 상태가 변하게 된다.

State	Next
000	0
001	1
010	0
011	1
100	0
101	1
110	0
111	1

(그림4)규정표



(그림5)세포상태변화

3) 환자와 대조군을 판단하는 기준은 결과 세포 중 하나를 선택하여 그 값이 1인지 0인지에 따라서 결정하게 된다. 이때 1/0에 환자와 대조군을 할당하게 된다.

2.2 유전 알고리즘(Genetic Algorithm)

2.2.1 일반적인 유전 알고리즘

유전 알고리즘은 원하는 해답을 얻기 위한 다항 시간(polynomial time) 알고리즘이 존재하지 않을 때 근사 해를 얻기 위한 알고리즘이다. 이 연구에서 최적의 CA를 얻기 위한 방법으로 유전 알고리즘을 사용하였다.

2.2.2 유전 알고리즘의 변형

이 연구에서 CA는 각 성분마다 가지는 특성들이 다르고 그 값의 범위도 다르기 때문에 일반적인 유전적 알고리즘을 적용하기 어렵고 지역적 근사해(local optima)의 문제가 존재하기 때문에 유전 알고리즘을 변형하여 사용하였다.

한 세대(population)는 우수 개체(Top CA)와 일반적인 개체(CA individuals)로 구성되고 우수 개체는 세대를 거치면서 우수한 성능(fitness)을 가지는 개체들을 모아 놓은 것이다.

한 세대에서 다음 세대로 바뀔 때, 먼저 일반 개체의 모든 속성들이 랜덤하게 다시 생성되고, 우수 개체의 속성들이 일반 개체의 속성들로 유전되고, 생성된 세대의 성능이 계산되고, 계산된 성능에 따라서 우수 개체 보다 뛰어난 성능의 일반 개체가 나타나면 우수 개체 중 가장 성능이 낮은 개체를 대신하게 된다.

각 개체들의 성능을 결정하는 함수는 민감도(Sensitivity)와 특이도(Specificity)의 함수로 결정된다.

※ *성능 = 민감도^2 * 특이도*

성능을 위의 식과 같이 결정한 이유는 일반적인 진단 모델에서 민감도가 특이도 보다 중요하다고 생각되었고, 실제 자료로 연구를 해 보았을 때 Moore의 연구에서와 같이 정확도를 성능 함수로 사용하였을 때는 민감도가 현저하게 낮은 경우가 발생하였다.

2.3 사용한 자료

149명의 aspirin-induced asthma (AIA), 181명의 aspirin-tolerant asthma (ATA), 153 명의 대조군 자료가 사용되었다. 자료들은 각 개인이 49개의 SNP 자료로 구성되어 있고 이 자료는 우성/우성, 우성/열성, 열성/열성, 미기재(missing)되어 있다. 실제 자료에서 미기재가 있는 자료에 대해서 전처리 과정을 거쳐서 미기재 자료가 없는 자료로 만드는 방법도 있지만 미기재가 포함된 자료도 의미를 가지는 것이라 생각되기 때문에 포함하고 연구를 진행하였다.

2.4 실험 결과

자료에서 미기재 되어 있는 자료에 대한 SNP에 대해서 일정 확률로 발생시키거나 SNP들의 연결 상관 관계 (linkage disequilibrium)에 따라서 채워 넣는 방법들이 존재하나, 잘못 채워질 확률이 존재하므로, 이 연구에서 CA에 선택되어진 SNP들의 조합 중 미기재가 포함되어 있는 경우에만 자료를 배제하였다.

2.4.1 AIA 대 ATA

AIA를 환자군으로 ATA를 대조군으로 실험 했을 때는 SNP 9,2,28,8번이 선택되었고 이때의 판단 정확도(accuracy)는 0.7391 민감도(sensitivity)는 0.893 특이도(specificity)는 0.514이고 분할표는 <표1>과 같다.

<표1> 분할표

	Test AIA	Test ATA
Real AIA	67	8
Real ATA	17	18

2.4.2 AIA 대 NC

AIA를 환자군으로 NC를 대조군으로 실험하였을 때, SNP는 7, 44, 24, 28이 선택되었고 이때 판단 정확도는 0.7273, 민감도는 0.8658, 특이도는 0.5000 이고 자료의 분할표는 <표2>와 같다.

<표2> 분할표

	Test AIA	Test NC
Real AIA	71	11
Real NC	25	25

2.4.3 ATA 대 NC

ATA를 환자군으로 NC를 대조군으로 실험하였을 때, SNP는 28, 7, 8, 9가 선택되었고, 이때 판단 정확도는 0.6941, 민감도는 0.7893, 특이도는 0.625, 자료의 분할표는 <표3>와 같다.

<표3> 분할표

	Test Yes	Test No
Real ATA	29	8
Real NC	18	30

2.4.4 분석

앞의 결과를 보면 자료들이 미기재 되어 있는 경우가 많은 SNP가 선택되어 지는 경향이 발생함을 알 수 있다. 실제 ATA와 AIA의 합이 330명인데 110명만 선택되는 경우가 발생함을 알 수 있다.

3. 결론

천식과 같은 복합질환의 작용 모델을 찾고 그 원인이 되는 SNP를 찾기 위하여 유전 알고리즘과 셀룰러 오토마타를 사용하였다.

Moore 등의 이전 연구에서는 특정한 규칙을 따르는 프로그램에 의해 생성된 자료를 가지고 실험을 해 보았으나, 본 연구에서는 실제 AIA환자와 ATA 환자, 대조군을 각각 AIA 대 ATA, AIA 대 대조군, ATA 대 대조군으로 실험하였다.

위와 같은 방법으로 실험을 한 결과 미기재 자료가 많은 SNP로 치우치는 현상이 발생하여, 미기재를 많이 포함하는 자료가 불이익을 받도록 유전 알고리즘의 성능함수를 개선하여 실험하여 보는 것이 향후 연구 과제이다.

참고 문헌

1. D.W. Hosmer and S. Lemeshow, "Applied Logistic Regression", John Wiley & Sons, New York, 2000.
2. S. Papadokonstantakis , A. Lygeros, S. Jacobsson, "Comparison of Recent Methods for Inference of Variable Influence in Neural Networks" Neural Net. 19(4):500-513. 2006.
3. Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, H. Honda, "Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma" BMC Bioinformatics 5:120, 2004.
4. A. G. Heidema, J. M. Boer, N. Nagelkerke, E.C. Mariman, D. L. van der A, E. J. Feskens, "The Challenge for Genetic Epidemiologists : How to Analyze Large Numbers of SNPs in Relation to Complex Disease", BMC Genetics 7(23), 2006.
5. K. Miyak, K. Omae, M. Murata, N. Tanahashi, I. Saito, K. Watanabe, "High Throughput Multiple Combination Extraction from Large Scale Polymorphism Data by Exact Tree Method", J. Hum. Genet. 49(9):455-462, 2004.
6. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.
7. J. Moore, L. Hahn, M. Ritchie, "Power of Multifactor Dimensionality Reduction for Detecting Gene-gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity", Genet. Epidem. 24(2):150-157, 2003.
8. J. Moore, L. Hahn, "A Cellular Automata Approach to Detecting Interactions Among Single-Nucleotide Polymorphisms in Complex Multifactorial Diseases", Pac. Symp. Biocomput. pp. 53-64, 2002 .