

엔트로피 거리와 SVM를 이용한 SNP 군집분석과 천식 유형 예측

이 중 섭[†] · 신 기 섭^{††} · 위 규 범^{†††}

요 약

단일염기다형성은 인간 게놈 구조 연구의 중요한 도구이다. 대량의 유전자 표현형 데이터에서의 군집 분석은 생물학적으로 연관이 있는 유전자 군을 발견하거나 유전자간 상호작용 네트워크를 생성하는데 유용하다. 본 논문에서는 엔트로피 거리를 기반으로 계층적 군집 분석 방법을 사용하여 천식환자군과 정상대조군의 군집을 형성하고 비교하였고 5개짜리 군집에서 두 군의 의미 있는 차이점이 나타남을 보였다. 천식환자군의 각 군집에서의 대표 SNP들의 조합의 질병 예측 정확도를 지지벡터기계를 이용하여 측정하여, 천식의 두 유형을 진단할 수 있는 최상의 조합을 찾았다. 최상의 조합은 유전자 ALOX12에 있는 단일염기다형성을 포함한 5개로 구성된 모델이며 66.41%의 아스피린 내성 천식 질병에 대한 예측 정확도를 갖는다.

키워드 : 천식, 단일염기다형성, 엔트로피 거리, 군집분석

Cluster Analysis of SNPs with Entropy Distance and Prediction of Asthma Type Using SVM

Jungseob Lee[†] · Kiseob Shin^{††} · Kyubum Wee^{†††}

ABSTRACT

Single nucleotide polymorphisms (SNPs) are a very important tool for the study of human genome structure. Cluster analysis of the large amount of gene expression data is useful for identifying biologically relevant groups of genes and for generating networks of gene-gene interactions. In this paper we compared the clusters of SNPs within asthma group and normal control group obtained by using hierarchical cluster analysis method with entropy distance. It appears that the 5-cluster collections of the two groups are significantly different. We searched the best set of SNPs that are useful for diagnosing the two types of asthma using representative SNPs of the clusters of the asthma group. Here support vector machines are used to evaluate the prediction accuracy of the selected combinations. The best combination model turns out to be the five-locus SNPs including one on the gene ALOX12 and their accuracy in predicting aspirin intolerant asthma disease risk among asthmatic patients is 66.41%.

Keywords : Asthma, Single Nucleotide Polymorphism, Entropy Distance, Cluster Analysis

1. 서 론

단일염기다형성(SNP)은 인간의 게놈(genome) 구조를 연구하는데 있어 중요한 도구로 알려져 있다. SNP는 인간 유전체를 이루는 전체 30억 염기 중 0.1%정도에 해당되는 부분으로 개인 간 형질의 차이를 결정하는 염기서열이다. 현재

질병의 표현형과 SNP 유전형과의 연관성에 대한 연구가 활발하게 진행되고 있으며, 이미 천식, 암 등 여러 가지 복합 질환과 관련된 연구 성과가 나타나고 있다[1, 2, 3, 5]. 이러한 SNP와 질병과의 관계에 대한 연구 결과물이 축적될 경우 이것과 개인에 대한 SNP 정보를 비교분석함으로써 그 개인이 특정질환에 대해 어느 정도의 감수성이 있는지를 알 수 있고, 질병의 예방과 치료에도 활용될 수 있게 될 것이다.

질병과 SNP의 연관성에 대한 지금까지의 연구 방법의 가장 큰 문제점은 시간 복잡도가 크다는 것이다. 특히 SNP 칩과 같은 대용량의 자료의 분석에는 어려움이 예상된다. 질병과 대량의 SNP 사이의 연관성을 직접 분석하는 대신에, SNP들을 유사한 군집으로 먼저 분류한 후 질병과의 연

※ 이 논문은 2006년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2006-000-10775-0).
† 정 회 원: 아주대학교 자연과학부 교수
†† 준 회 원: 아주대학교 수학과 박사과정
††† 정 회 원: 아주대학교 정보컴퓨터공학부 교수
논문접수: 2010년 10월 22일
수정일: 1차 2010년 12월 22일, 2차 2011년 2월 21일
심사완료: 2011년 2월 21일

관성을 연구하는 간접적인 방법을 시도할 필요가 있다. 본 논문에서 처음으로 시도되는 이 방법은 시간 복잡도가 작지만 비교적 높은 질병 예측정확도를 제공한다.

분석 대상 개체 사이의 거리를 무엇으로 선택하는가하는 것은 군집 분석의 관건이다. 상호정보(mutual information)는 RNA의 군집 분석에 활용된 바 있지만[4], 이것은 수학적 측도(metric)가 아니므로 이를 이용하여 형성한 군집들의 차별성을 보장하지 못한다. 본 논문에서는 상호정보를 대신하여 정보이론에서 알려진 엔트로피 거리를 이용하여 군집 분석을 실시한다. 엔트로피 거리가 측도가 된다는 것은 알려진 사실이다[7].

<표 1> 천식과 관련된 SNPs

ID	SNP symbol
SNP1	ALOX5_p1_1708_G>A
SNP2	ALOX5_q3_1728_A>G
SNP3	ALOX12_p1_822C>G
SNP4	ALOX12_p2_540G>A
SNP5	ALOX12_q1_261_G>A
SNP6	ALOX12_q2_322_A>G
SNP7	ALOX15_P_272_C>A
SNP8	B2ADR_q1_46_A>G
SNP9	B2ADR_q2_79_C>G
SNP10	CCR3_p2_520_T>G
SNP11	CCR3_p3_174_C>T
SNP12	CCR3_p1_22557_G>A
SNP13	FCER1A_p1_344_C>T
SNP14	FCER1A_p2_95T>C
SNP15	FCER1B_p1_109_T>C
SNP16	FCER1B_q1_237_A>G
SNP17	FCER1G_p1_237A>G
SNP18	FCER1G_p2_54G>T
SNP19	IL10_p1_1082_A>G
SNP20	IL10_p2_819_T>C
SNP21	IL13_p1_1510_A>C
SNP22	IL13_p2_1055_C>T
SNP23	IL13_q_110_G_A
SNP24	IL18_p2_607_A>C
SNP25	IL18_p3_137_G>C
SNP26	IL18_q5_13881_A>C
SNP27	LTC4S_p2_444_A>C
SNP28	TBXA2R_q1_795_T>C
SNP29	TGFB_p1_509_C>T
SNP30	TNFA_p1_1031_T>C
SNP31	TNFA_p2_863_C>A
SNP32	TNFA_p3_857_C>T
SNP33	TNFA_p5_308_G>A
SNP34	TNFA_p6_238_G>A

본 논문의 주요 내용을 요약한다. 첫째, 천식환자군과 정상대조군 내의 모든 SNP 사이의 엔트로피 거리를 계산하여 유사행렬(similarity matrix)을 만들고 평균결합군집(average linkage clustering) 방법을 이용하여 SNP 사이의 상호작용 수형도를 건설한다. 천식환자군과 정상대조군에서 형성된 SNP 군집의 차이를 비교한다. 둘째, 천식환자군에 대한 군

집분석의 결과를 이용하여, 천식의 두 가지 유형(아스피린 내성 천식과 아스피린 불내성 천식)을 진단하는 유용한 SNP조합을 찾는다. 같은 군집에 속하는 SNP는 유사한 유전형질을 갖는다고 생각할 수 있으므로, 천식환자군에서 형성된 각 군집을 대표하는 SNP들을 결정한다. 이들로 구성된 SNP 조합의 아스피린 내성 천식 진단 예측정확도를 지지벡터기계(SVM) 방법을 이용하여 측정한다. 마지막으로 질병과 SNP 사이의 연관성에 관한 기존의 연구 방법들과 본 논문의 방법을 비교한다.

2. 데이터

본 연구는 소속병원 알레르기-류마티스내과 외래를 내원한 256명의 천식환자군과 145명의 정상대조군을 대상으로 하였으며, 천식환자군은 병력과 아스피린 경구 유발 검사 결과를 바탕으로 97명의 아스피린 불내성 천식(AIA)군과 159명의 아스피린 내성 천식(ATA) 군으로 분류하였다. 본 연구에 참여한 모든 환자는 병원의 임상시험심의위원회(IRB)에서 승인한 바에 따라 고지 후 동의서를 작성하였고 유전자 정보는 코드화하여 분석에 활용되었다. 분석에 사용한 SNP들은 일반 천식과 관련이 있다고 알려진 34개의 SNP를 대상으로 하였다[2, 6]. <표 1>은 실험에 사용된 SNP에 대한 정보이다.

3. 방법

3.1 엔트로피 거리

유한 집합 S 에 대한 분할 $X = \{X_1, X_2, \dots, X_n\}$ 에 대하여 X 의 엔트로피는

$$H(X) = - \sum_{i=1}^n \frac{|X_i|}{|S|} \log \frac{|X_i|}{|S|}$$

이고 여기서 절댓값 기호는 집합의 원소의 개수를 나타낸다. 같은 집합 S 에 대한 다른 분할 $Y = \{Y_1, Y_2, \dots, Y_m\}$ 이 주어지면

$$H(X, Y) = - \sum_{i=1}^n \sum_{j=1}^m \frac{|X_i \cap Y_j|}{|S|} \log \frac{|X_i \cap Y_j|}{|S|}$$

을 X 와 Y 의 결합 엔트로피라 한다. 분할 X 와 Y 의 엔트로피 거리(Entropy Distance)를

$$ED(X, Y) = 2H(X, Y) - H(X) - H(Y)$$

로 정의한다[7]. 그러면 $ED(X, Y)$ 는 측도(metric)의 공리를 만족함을 증명할 수 있다. 즉, X, Y, Z 가 집합 S 위의 분할이면 다음이 성립한다.

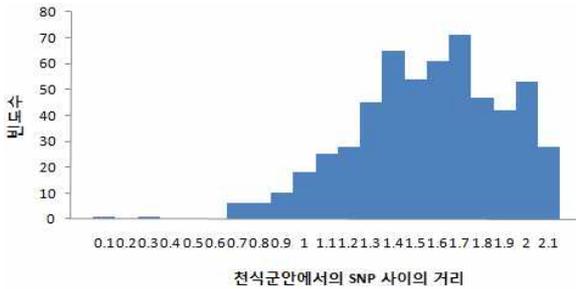
- (a) $ED(X, Y) \geq 0$;
- (b) $ED(X, Y) = 0$ 일 필요충분조건은 $X = Y$;

- (c) $ED(X, Y) = ED(Y, X)$;
- (d) $ED(X, Y) \leq ED(X, Z) + ED(Z, Y)$.

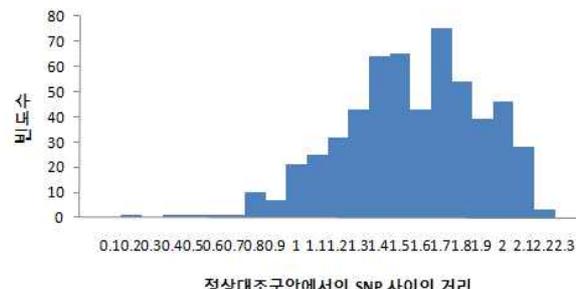
이것은 $ED(X, Y)$ 가 직관적인 거리 개념과 일치함을 보여준다.

본 논문에서는 전체집합 S 를 천식환자군 또는 정상대조군으로 설정하고 각 SNP의 유전형에 의한 분할을 생각한다. 두 개의 SNP에 의한 분할 사이의 엔트로피 거리는 두 SNP의 유사성을 측정하는 척도로서, 엔트로피 거리가 작을수록 연관성이 커진다. 전체집합이 정상대조군일 때 34개의 SNP 사이의 엔트로피 거리를 모두 계산하고, 전체집합이 천식환자군일 때와 비교하고, 정상대조군에서 유사성이 높은 SNP들의 군집과 천식환자군에서 유사성이 높은 SNP들의 군집을 찾는데 사용된다.

엔트로피 거리는 성질 (d)를 만족시키므로 SNP 사이의 관계를 보다 명확히 할 수 있다는 장점이 있다. 예를 들면 SNP1과 SNP2의 거리와 SNP2와 SNP3의 거리를 안다면 SNP1과 SNP3의 거리는 SNP1과 SNP2의 거리와 SNP2와 SNP3의 거리의 합보다 작거나 같다는 사실을 알 수 있다.



(평균 거리 : 1.424, 표준편차 : 0.477)



(평균 거리 : 1.419, 표준편차 : 0.482)

(그림 1) 천식환자군과 정상대조군내에서의 SNP 사이의 거리 분포 비교

3.2 계층적 군집 분석 방법

군집분석이란 다양한 특성을 지닌 대상들을 유사성에 기초하여 동질적인 집단으로 분류하는데 이용되는 기법을 말한다. 본 논문에서는 계층적 군집 분석 방법(hierarchical clustering) 중 하나인 평균 결합 군집 방법(average linkage clustering method)[8]을 사용하여 천식환자군과 정상대조군

에 대하여 따로따로 군집분석을 실시하였다.

구체적인 방법의 단계는 다음과 같다. 먼저 엔트로피 거리를 사용하여 각 군에서 34개의 SNP 사이의 거리 ${}_{34}C_2 = 561$ 개를 계산하고 이를 바탕으로 SNP에 대한 유사행렬을 만든다. 먼저 1개의 SNP로 이루어진 34개의 군집에서 출발하여 거리가 가장 가까운 두 군집을 결합하여 33개의 군집을 결정한다. 33개의 군집에서 거리가 가장 가까운 두 군집을 결합하면 32개의 군집을 얻고, 이 과정을 계속하여 1개의 군집(즉, 전체 군)에 이를 때까지 수행한다. 여기서 두 군집 사이의 거리는 각각의 군집내의 모든 SNP들 간의 평균거리를 의미한다. 즉, 다음과 같은 식으로 표현된다.

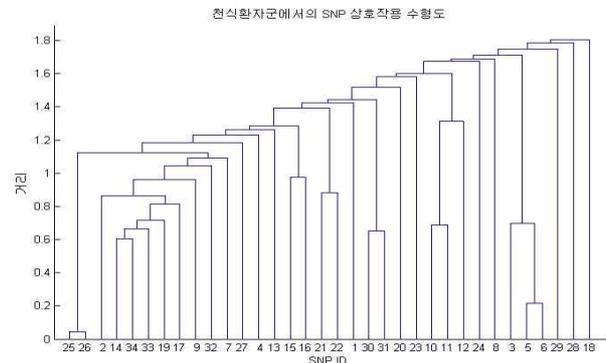
$$d(C_i, C_j) = \frac{1}{m_i m_j} \sum_{x \in C_i, y \in C_j} d(x, y)$$

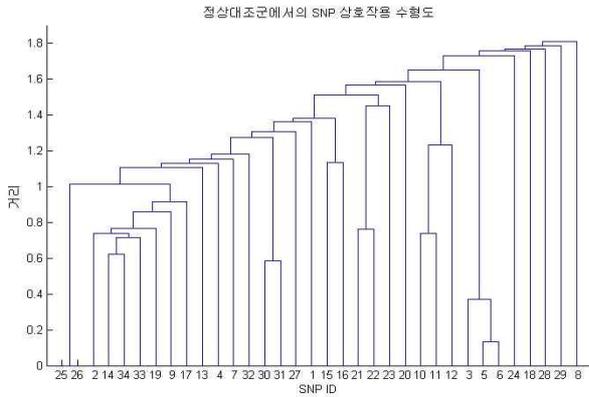
여기서 d 는 거리를 의미하고 C_i 와 C_j 는 각각 크기 m_i 와 m_j 를 가지는 군집이다.

4. 천식환자군과 정상대조군의 군집결과 비교

3절의 방법을 이용하여 천식환자군과 정상대조군에서 각각 군집을 형성하고 비교한 결과를 설명한다. 먼저, 천식환자군과 정상대조군 내의 모든 SNP 사이의 엔트로피 거리를 계산하였다. 엔트로피 거리의 평균과 표준편차 등은 두 군에서 큰 차이가 없으나, 분포는 다소의 차이가 있고 이것은 두 군의 군집 형성에 영향을 줄 것으로 보인다(그림 1). 이제 엔트로피 거리로 구성된 유사행렬을 만들고, 평균 결합 군집 방법을 적용하여 두 군의 34개짜리부터 1개짜리 군집까지 형성하였다. 이 군집 분석의 결과는 두 군의 SNP들의 상호작용 수형도로 나타난다(그림 2). 그리고 두 군에서 형성된 군집들의 개수가 같을 때를 비교하였다.

군집의 개수가 1~2개일 때는 의미가 없고, 군집의 개수가 7개 이상인 경우는 군집의 개수가 6개인 경우와 결과가 크게 다르지 않으므로 그 결과를 생략하였다. 두 군의 3, 4, 5, 6개짜리 군집 형성 결과를 <표 2>와 <표 3>에 제시하였다. <표 2>와 <표 3>에서의 숫자는 <표 1>에 제시한 SNP 목록의 ID를 의미한다.





(그림 2) 천식환자군과 정상대조군의 SNP 상호작용 수형도의 비교

두 군의 3, 4개짜리 군집을 비교해 보면, 천식환자군에서는 SNP 18로 이루어진 군집이 나타나는 반면에 정상대조군에서는 SNP 8로 이루어진 군집이 나타나고 있다. 두 군의 차이는 5, 6개짜리 군집을 형성할 때 가장 크게 나타난다. 천식환자군의 5, 6개짜리 군집 중에는, 정상대조군에는 나타나지 않는 SNP 3, SNP 5, SNP 6으로 구성된 군집이 확인된다. 3, 4, 5, 6개짜리 군집에서 두 군의 군집이 가장 큰 차이를 보이므로 여기서 두 유형의 천식 진단에 유용한 SNP의 조합을 찾을 수 있음을 예측할 수 있다.

SNP 3, SNP 5, SNP 6은 모두 같은 유전자에 위치하는 SNP로서 정상대조군에서보다 천식환자군에서 더 강한 연관성을 보여 주고 있다. 이에 대한 의학적, 생물학적인 해석은 추후 연구가 필요한 부분이다.

<표 2> 천식환자군에서의 3, 4, 5, 6개짜리 군집 (대표 SNP는 진하게 표시)

3개짜리 군집	4개짜리 군집	5개짜리 군집	6개짜리 군집
SNP 18	SNP 18	SNP 18	SNP 18
SNP 28	SNP 28	SNP 28	SNP 28
SNP 13를 비롯한 32개의 SNP	SNP 29	SNP 29	SNP 29
	SNP 22를 비롯한 31개의 SNP	SNP 3, 5, 6	SNP 3, 5, 6
		SNP 13를 비롯한 28개의 SNP	SNP 8
			SNP 13 를 비롯한 27개의 SNP

<표 3> 정상대조군에서의 3, 4, 5, 6개짜리 군집

3개짜리 군집	4개짜리 군집	5개짜리 군집	6개짜리 군집
SNP 8	SNP 8	SNP 8	SNP 8
SNP 29	SNP 28	SNP 28	SNP 28
SNP 14를 비롯한 32개의 SNP	SNP 29	SNP 29	SNP 29
	SNP 19를 비롯한 31개의 SNP	SNP 18	SNP 18
		SNP 25를 비롯한 30개의 SNP	SNP 24
			SNP 25 를 비롯한 29개의 SNP

5. 천식유형의 진단

5.1 군집을 대표하는 SNP의 결정

같은 군집에 있는 SNP들은 유사한 유전형 분포를 갖기 때문에, 각 군집을 대표하는 SNP를 결정하여 질병 진단에 활용하는 것이 편리할 것이다. 군집들 사이의 거리를 가장 잘 표현하는 SNP들을 군집의 대표로 정하는 것이 자연스럽다. k 개의 군집 C_1, C_2, \dots, C_k 가 주어졌을 때, 각 군집의 대표 SNP는 아래의 식 (1)을 만족하는 $x_1 \in C_1, x_2 \in C_2, \dots, x_k \in C_k$ 로 정의한다.

$$\min_{x_1 \in C_1, x_2 \in C_2, \dots, x_k \in C_k} \sum_{i,j=1}^k |d(x_i, x_j) - d(C_i, C_j)| \quad (1)$$

여기서 $d(C_i, C_j)$ 는 군집간의 평균거리이다.

5.2 진단정확도

아스피린 내성에 따른 천식의 두 가지 유형의 진단에 유용한 SNP의 조합을 찾기 위하여, 천식환자군에서 형성된 군집에 대해서 식 (1)을 사용하여 대표 SNP들을 추출하였다(<표 2>에 진하게 표시). 예를 들면 천식환자군의 3개짜리 군집의 대표자는 각각 SNP 13, SNP 18, SNP 28이며, 이들 SNP 조합의 천식 유형에 대한 예측정확도는 57.42%이다(<표 4>). 여기서 예측정확도는 방사 기저 함수(RBF : Radial Basis Function)를 사용한 지지벡터기계(SVM)[9] 방법을 통하여 계산하였고 leave-one-out 교차검증을 적용하였다.

여기에서 지지벡터기계를 어떻게 적용했는지 구체적으로 설명한다. 군집 분석 결과에 의해 SNP 13, SNP 18, SNP 28이 얻어졌으므로 이제 이들 SNP의 유전형에 따라 아스피린 불내성 천식과 아스피린 내성 천식을 진단해야 한다. 256명의 천식환자에 대한 데이터를 다음과 같은 방법으로 변환한다. 두 유형의 천식을 구분하기 위해 아스피린 불내성 천식환자는 라벨 1로 아스피린 내성 천식환자는 라벨 -1로 설정한다. 환자의 SNP 유전형이 major allele이면 1의 값을 minor allele이면 -1의 값을 할당한다. SNP 13의 major allele은 C이고 minor allele은 T이다. 따라서 환자의 SNP 13의 유전형이 CC이면 (1, 1), CT이면 (1, -1), TT이면 (-1, -1)의 좌표에 대응된다. 이러한 방법으로 환자의 SNP 13, SNP 18, SNP 28의 유전형을 동시에 고려하면 6차원 좌표에 대응된다. 이제 변환된 데이터에 지지벡터기계를 적용한다.

동일한 방법으로 4, 5, 6개짜리 군집의 대표 SNP를 결정하여 이들 조합의 예측 정확도를 계산하였다(표4). 5개짜리 군집의 대표인 SNP 3, SNP 13, SNP 18, SNP 28, SNP 29로 구성된 조합에서 최고의 예측정확도 66.41%를 얻었다.

최상의 조합을 구성하는 SNP들이 위치한 5개의 유전자 중에서 ALOX12, TBXA2R, TGFβ는 천식의 병인기전에 관

런된 염증 매개 물질에 의한 만성적 염증에 중요한 기능을 담당하고, FCER1A와 FCER1G는 면역글로불린을 통한 면역 반응의 증가에 중요한 기능을 담당하고 있다. 본 연구의 분석 결과를 통해 산출된 5개의 유전자간의 상호작용에 대한 추가적인 연구가 필요하겠다.

5.3 다른 방법과의 비교

참고문헌 [5]에서 제시한 방법으로 N 개의 SNP로부터 질병에 관련된 SNP를 찾으려면 2^N 개의 모든 SNP조합에 대하여 상호정보량을 계산해야 한다. 본 논문의 방법에서는 $\frac{N(N-1)}{2}$ 개의 엔트로피 거리를 계산하고 군집 분석을 해야 한다. 군집분석의 계산복잡도는 $O(N^3)$ 이라는 것이 알려져 있다[11]. 따라서 SNP칩과 같은 대용량 데이터 분석에는 참고문헌 [5]에서 제시한 방법보다 본 논문에서 제시한 방법이 더 효율적이다.

실제로 참고문헌 [5]에서 제시한 방법을 사용했을 경우에 ${}_{34}C_5$ 개의 상호정보량의 계산에 소요되는 CPU-시간은 6.1214×10^3 초인 반면에, 본 논문에서 제시하는 방법을 사용하여 34개 SNP 사이의 엔트로피 거리 계산 및 군집분석에 소요된 CPU-시간은 1.0075초이다. CPU-시간을 계산하는데 사용된 환경은 Intel Core2 CPU 2.13GH의 CPU와 2GB의 DDR2램을 장착한 컴퓨터이며 Matlab으로 구현되었다.

동일한 SNP를 이용하여 참고문헌 [5]에서 제시된 정규 상호정보 방법으로 찾아낸 모델의 예측정확도는 64.84% 이고 참고문헌 [2, 10]에서 제시된 MDR (Multifactor Dimensionality Reduction) 방법으로 찾아낸 모델의 예측정확도는 64.48%로서, 본 논문에서 제시한 방법의 예측정확도가 다소 높다.

〈표 4〉 SNP 조합의 예측정확도 비교

조합의 개수	SNP 조합	예측정확도(%)
3	SNP 13,18,28	57.42%
4	SNP 18,22,28,29	62.10%
5	SNP 3,13,18, 28,29	66.41%
6	SNP 3,8,13,18,28,29	63.69%

$$* \text{예측정확도} = \frac{(TP+TN)}{(TP+TN+FN+FP)} \times 100$$

TP : 진양성(True Positive)

TN : 진음성(True Negative)

FP : 위양성(False Positive)

FN : 위음성(False Negative)

6. 결 론

복합질환에 관련된 SNP를 찾아서 진단 마커로 활용하는 다양한 방법들이 개발되고 있다. 본 논문에서는 최초로 엔트로피 거리를 이용하여 천식 관련 SNP에 대한 계층적 군집분석을 실시하였다. 그리고 찾아낸 군집들의 대표 SNP로 구성된 조합이 질병 유형 진단에 사용될 수 있음을 보였다. 이 방법은 질병에 관련된 SNP들의 조합을 직접 찾는 방법에 비해 계산 복잡도가 낮으면서도 질병 예측 정확도는 높은 것으로 나타났다. 본 논문에서 제시된 방법은 대용량의 유전 데이터를 분석하는데 활용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] D. H. Kim, K. B. Ham, J. Kim, "Effective Analysis of Related Chronic Hepatitis Using SNP", Proc. of the KIISE Korea Computer Congress 2006, Vol.33, No.1(A), pp.19-21, 2006. (in Korean)

[2] S. H. Kim, H. H. Jeong, B. Y. Cho, et al, "Association of four-locus gene interaction with aspirin-intolerant asthma in Korean asthmatics", J. Clin. Immunol., Vol.4, No.4, pp.336-342, 2008.

[3] E. G. Julie, E. M. Leah, T. L. Brian, et al, "Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis", Int. J. Cancer, Vol.118, No.7, pp.1790-1797, 2006.

[4] A. J. Butte, and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements", Pac. Symp. Biocomput., Vol.5, pp.415-426, 2000.

[5] J. S. Lee, S. H. Kim, K. S. Shin, et al, "Analysis of Asthma Related SNP Genotype Data Using Normalized Mutual Information and Support Vector Machines", Journal of KIISE : Software and Applications, Vol.36, No.9, pp.691-696, 2009. (in Korean)

[6] S. H. Kim, H. S. Park, J. W. Holloway, et al. "Association between a TGF β_1 promoter polymorphism and rhinosinusitis in aspirin intolerant asthmatic patients", Respir Med. Vol.101, pp.490 - 495, 2007.

[7] T. M. Cover, and J. A. Thomas, Elements of Information Theory, 2nd Ed., Wiley, 2006.

[8] M. B. Eisen, T. P. Spellman, P. O. Brown, et al, "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci., Vol.95, No.25, pp.14863-14868, 1998.

[9] <http://svmlight.joachims.org/>.

[10] J. H. Moore, J. C. Gilbert, C. T. Tsai, et al, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studied of human disease susceptibility", J. Theor. Biol., Vol.241, pp.252-261, 2006.

[11] P. N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining", Pearson Education, 2006.



이 중 섭

e-mail : jslee@ajou.ac.kr
1976년 서울대학교 수학과(학사)
1982년 서울대학교 수학과(석사)
1989년 University of Michigan(박사)
1992년~현 재 아주대학교 자연과학부
교수
관심분야: 동역학계



위 규 범

e-mail : kbwee@ajou.kr
1978년 서울대학교 수학과(학사)
1985년 University of Wisconsin
전산학과(석사)
1992년 Indiana University(박사)
1993년~현 재 아주대학교 정보컴퓨터
공학부 교수
관심분야: 컴퓨팅이론, 생물정보학



신 기 섭

e-mail : k1430@ajou.ac.kr
2000년 청주대학교 수학교육학과(학사)
2002년 인하대학교 수학과(석사)
2003년~현 재 아주대학교 수학과 박사
과정
관심분야: 동역학계, 생물정보학