*The Korean Journal of Applied Statistics* (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.563
**25**(4), 563–573

# Detecting Genetic Association and Gene-Gene Interaction using Network Analysis in Case-Control Study

Seohoon Jin[1] · Minhee Lee[2] · Hyo-Jung Lee[3] · Mira Park[4]

[1]Department of Informational Statistics, Korea University
[2]Department of Statistics, Korea University; [3]Department of Statistics, Korea University
[4]Department of Preventive Medicine, Eulji University

(Received April 10, 2012; Revised May 2, 2012; Accepted June 5, 2012)

**Abstract**

Various methods of analysis have been proposed to understand the gene-disease relation and gene-gene interaction effect for a disease through comparison of genotype in case-control study. In this study, we proposed the method to detect a genetic association and gene-gene interaction through the use of a network graph and centrality measures that are used in social network analysis. The applicability of the proposed method was studied through an analysis of real genetic data.

Keywords: Network analysis, genetic association, gene-gene interaction, centrality measures.

## 1. 서론

최근 질병이나 특정 형질에 관한 유전적 영향을 조사하여 질병의 진단과 예방 및 치료를 개선하기 위한 연구가 활발히 진행되고 있다. 유전적 연관분석은 집단기반연구와 가족기반연구로 나눌 수 있다. 집단기반연구는 주로 사례-대조 설계로 이루어지며, 관심 있는 질병을 확진 받은 환자집단인 사례군과 모집단에서 랜덤하게 선택되고 관심질병이 없는 대조군에서의 대립형질이나 유전자형 빈도를 비교하게 된다. 이러한 설계는 가족연구에 비해 비용이 적게 들고 CDCV(common disease, common variants)연구에 대해 검정력이 있다고 알려져 있다 (Balding, 2006).

유전자형과 질병간의 관계를 파악하기 위한 유전적 연관분석(association analysis)이나 유전자간의 상호작용이 질병에 미치는 영향을 파악하기 위한 유전자-유전자간 상호작용분석(gene-gene interaction analysis)을 위하여 CMH(Cochran-Mantel-Haenszel) 검정이나 로지스틱회귀분석, MDR(multidimensional reduction), random forest 등의 여러 통계적 방법이 적용되고 있다 (Lee 등, 2010; Jung 등, 2011; Balding, 2006; Cordell, 2009; Ritchie 등, 2001). 본 연구에서는 네트워크방식을 이용하여 질병

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (2011-0004376).
[4]Corresponding author: Professor, Department of Preventive Medicine, Eulji University, Daejeon 301-832, Korea. E-mail: mira@eulji.ac.kr

간의 관계 및 유전자간의 상호작용과 질병과의 연관성을 탐색하는 방법을 모색하였다. 다수의 점과 그 점들을 연결하는 선으로 구성된 망, 즉 복잡한 네트워크는 사회학, 컴퓨터공학을 비롯하여 생물학적 시스템을 설명하는데 사용될 수 있다. 유전학 분야에서도 네트워크를 이용한 분석이 종종 시도된 바 있다. Snel와 Huynen (2002)는 단백질들 사이의 상호작용의 결합을 이용, 단백질 상호작용 네트워크를 생성하고 분석하였다. 여기서는 네트워크 분석을 통해서 생물학적 과정 안에 포함되어있는 단백질 세트, 즉 기능 모듈을 탐색하였다. Yanai와 DeLisi (2002)는 기능적 연결을 이용한 네트워크 분석을 비교유전체의 한 방법으로 이용, 유전적 연관성을 분석하였다. Lindlof와 Olsson (2002)은 유전자 발현 데이터에서 유클리디안 거리, 피어슨 상관계수 그리고 상호 상관계수를 기반으로 한 네트워크를 추출, 각 네트워크를 분석한 결과를 비교하였다. 최근 Fortes 등 (2010)은 소의 유전자와 형질 데이터를 동시에 이용한 네트워크를 생성하여 분석한 바 있다.

여기서는 사회네트워크분석(social network analysis; SNA)에서 사용되는 기법들을 유전자 데이터에 적용하여 유전자와 질병간의 관계 및 유전자간의 상호작용과 질병과의 연관성을 파악하였다. 사회 네트워크는 다수의 개인이 일련의 관계에 의해 연결되는지를 파악하는 관계망으로서, 행위와 구조의 상호역동성을 설명하는 이론이다 (Sohn, 2002). 사회학에서 출발한 이 분석은 현재 경영, 금융, 통신을 비롯하여 생물학 분야까지 확대되어 널리 응용되고 있다.

사회네트워크분석에서는 네트워크내의 특정 노드의 중요도 또는 영향력을 수량화하기 위한 척도로서 중심성(centrality)지수를 사용하며, 이를 이용하여 네트워크내의 다양한 역할과 그룹화에 대한 정보를 제공한다. 즉, 연결자, 지도자, 고립자가 누구인지에 대한 정보와 어떤 그룹이 존재하고 그 그룹의 구성원은 누구이고, 그룹의 중심은 누구인지를 알아볼 수 있다. 또한 이러한 관계에 대한 시각적 분석을 가능하게 한다. 이러한 특성으로 인해, 유전자와 질환간의 연관분석에 있어 어떤 유전자가 이러한 역할을 하는지, 그리고 유전자 클러스터의 구성원과 중심은 어떤 유전자인지 알아보기 위한 방법으로 적절하다. 지금까지 생물학적 네트워크분석은 주로 단백질간 상호작용(PPI; protein-protein interaction), 유전자 조절 네트워크(GRN; gene regulatory network), 대사 네트워크(metabolic network) 등에 적용되었으며 (Zhu 등, 2007), 네트워크의 국부적 패턴인식에 초점을 맞추게 되는 경우가 많았다. 본 연구에서는 질병여부를 함께 분석한 네트워크를 통해 주효과를 검색하고, 사례군과 대조군을 별도 분석한 네트워크의 비교를 통해 유전자간 상호작용을 검색하며 각 상황에서 영향력이 있고 중심이 되는 유전자와 그들의 집단간 특성을 탐색하는 방안을 제시하였다. 2절에서는 사회네트워크 분석방법을 유전적 연관자료에 적용하기 위한 절차를 제안하였으며 3절에서 분석결과를 제시하였다. 4절에서 방법의 유용성과 한계에 대해 토의하였다.

## 2. 네트워크를 이용한 유전적 연관 및 상호작용분석방법

본 연구에서는 집단을 기반으로 하는 사례-대조군 연구에서 단일염기다형성(Single Nucleotide Polymorphism; SNP)의 유전자형을 이용하여 유전자-질병 및 유전자-유전자간 상호작용을 탐색하는 네트워크 방법을 고려하였다. 제안하는 분석절차는 다음과 같다.

[단계 1] 네트워크 데이터 생성

SNP간의 관계의 강도를 정의하는 가중 네트워크(weight network)를 구성한다. 각 원소로서 유전자형에서의 위험 대립형질(risk allele)의 수(0/1/2)를 이용하여 두 SNP간 비모수상관계수의 절대값을 구한

다. $n$개의 SNP이 있을 때 이는 다음과 같은 $n \times n$ 행렬로 표현된다.

$$\mathbf{G} = \begin{bmatrix} 0 & \cdots & g_{1j} & \cdots & \cdots & g_{1n} \\ \vdots & & \vdots & & & \vdots \\ \vdots & & 0 & & & \vdots \\ g_{i1} & \cdots & g_{ij} & 0 & \cdots & g_{in} \\ \vdots & & \vdots & & & \vdots \\ g_{n1} & \cdots & g_{nj} & \cdots & \cdots & 0 \end{bmatrix},$$

여기서 $g_{ij}$ $(\geq 0)$는 $i$번째 SNP과 $j$번째 SNP간의 비모수상관계수의 절대값을 의미한다.

[단계 2] 유전적 연관성 탐색

유전적 연관성 분석을 위해서는 가중 네트워크에 질병유무변수를 추가하여 $(n+1) \times (n+1)$ 가중네트워크를 생성한다. force-directed 알고리즘 (Fruchterman과 Reingold, 1991)이나 Kamada-Kawai 알고리즘 (Kamada와 Kawai, 1989) 등의 시각화 알고리즘을 이용하여 그래프로 표현한다. 이를 통해 질병에 대한 주효과를 탐색하고 서브그룹을 탐지한다. 본 연구에서 이용한 force-directed 알고리즘은 노드 간 연결선의 길이를 가능한 비슷하게 만들고 연결선의 교차를 최소화할 수 있도록 그림을 만들어 간다. 전체 시스템의 에너지를 최소화(힘의 균형화)하는 방향으로 노드를 움직여 재배치하는 것을 기본적 방법으로 하고 있다. 여기서 에너지는 두 노드 간 실제 거리와 그래프 상에서 두 노드 간 연결 길이의 차이로부터 계산된다. force-directed 알고리즘에서 노드와 연결선은 전기적 입자처럼 생각될 수 있는데 각 노드를 밀고 당기는 힘이 부여된다. 노드의 밀고 당기는 반복은 전체 노드가 평형상태에 다다를 때까지 진행된다 (Fruchterman과 Reingold, 1991).

[단계 3] 유전자-유전자간 상호작용 탐색

유전자-유전자간 상호작용 탐색을 위해서는 질병유무에 따라 두 개의 $n \times n$ 가중 네트워크를 생성한다. 시각화 알고리즘으로 구한 그래프를 통해 두 네트워크의 구조적 변화를 탐색한다.

[단계 4] 중심성 지수분석

각 SNP의 중심성 지수를 구하여 질병-유전자간 및 유전자-유전자간 관계를 확인한다. 상호작용탐색을 위한 네트워크에서는 두 네트워크에서 각 유전자들의 중심성 지수를 비교하여 네트워크와 각 네트워크 내의 유전자 특성을 비교한다. 또한 중심성 지수나 순위가 크게 변하는 유전자와 연결강도가 크게 변하는 연결선(link) 발견을 통해 특정 질환에 영향을 줄 가능성이 있는 유전자나 유전자 상호작용을 추측한다.

중심성 지수로서 다음과 같은 근접 중심성지수(closeness centrality index) $C_C(i)$와 중개 중심성지수(betweenness centrality index) $C_B(v)$를 사용한다 (Hanneman과 Riddle, 2005). 중심성지수의 계산을 위해서는 각 SNP간의 거리가 필요한데 여기서는 SNP간 상관계수의 역수를 거리로 이용하였다.

$$C_C(i) = \frac{\sum\limits_{j \neq i} \frac{1}{d(i,j)}}{n-1}, \quad i = 1, \ldots, n, \tag{2.1}$$

여기서 $d(i,j)$는 두 SNP간의 최단거리이다.

$$C_B(v) = \sum_{i \neq v} \sum_{j \neq v, \neq i} \frac{h_{ivj}}{h_{ij}}, \quad v = 1, \ldots, n, \tag{2.2}$$

여기서 $h_{ij}$는 SNP $i$에서 $j\,(\neq i)$로 가는 최단경로의 수이고, $h_{ivj}$는 SNP $i$에서 $j\,(\neq i, \neq v)$로 가는 최단경로 가운데 SNP $v\,(\neq i)$를 거치는 경로의 수이다. 이밖에도 연결선 수(degree centrality), 행렬의 고유값 분해를 이용한 고유벡터 중심성(eigenevector centrality) 등을 개별 개체의 중심성 정도를 나타내는 척도로 사용할 수 있다.

[단계 5] 대체 가능 SNP 탐색

구조적 동치의 관계거리를 이용하여 네트워크의 하부에 존재하는 부그룹(subgroup)을 탐색한다. 구조적으로 동일한 개체들은 상호 대체 가능하기 때문에 네트워크의 축소, 대체 SNP의 파악에 이용할 수 있다. 근사적 동치성을 포착하기 위하여 다음과 같이 개체 $i$와 $j$간 Hamming 거리를 정의한다 (Huh, 2010).

$$d_{ij}^H = \sum_{k \neq i} \sum_{k \neq j} (|g_{ik} - g_{jk}| + |g_{ki} - g_{kj}|). \tag{2.3}$$

이를 기반으로 군집분석을 실시하여 부그룹 및 대체 가능 SNP을 탐색한다. 여기서 $g_{ij}$는 분석에 이용한 가중네트워크 행렬 $G$의 $i$행 $j$열 원소이다.

## 3. 사례분석 결과

### 3.1. 데이터 및 분석방법

분석에 사용된 데이터는 원래 433명의 아토피성 피부염환자와 474명의 정상인을 대상으로 하여 5개의 유전자(IL5, IL8, IL5R, IL8RA, IL8RB)로부터 17개의 SNP을 타이핑한 것이다. 이들 SNP은 모두 하디-와인버그평형을 만족하였다. 이 SNP들은 Tagger program을 통해 LD(linkage disequilibrium) bin 방식으로 tagging SNP 중에서 결정되었으며, 각 bin에서 강한 LD($r^2 > 0.8$)를 보이는 SNP들 중에서 하나의 SNP만을 선택하는 방식으로 선정되었다 (Namkung 등, 2007). 여기서는 이 중에서 완전한 데이터를 가진 385명의 환자와 440명의 정상인에 대한 2개 유전자, 15개 SNP에 대한 유전자형자료가 사용되었다.

여기서는 네트워크 데이터 형성을 위해 비모수 상관계수 중 하나인 켄달의 타우-비($\tau_b$)의 절대값을 개체간의 관계의 가중치로 사용하여 가중 네트워크 분석을 실시하였다. 즉,

$$\tau_b = \frac{P - Q}{w_C w_R}, \tag{3.1}$$

여기서 $P$와 $Q$는 각각 일치쌍과 비일치쌍의 수이며, $w_C$는 서로 다른 열의 쌍의 수, $w_R$은 서로 다른 행의 쌍의 수이다. 상관계수값이 0.05보다 작은 경우 값을 0으로 바꾸었다. 네트워크 그래프를 그리기 위해서 R의 sna 프로시저를 이용하였다 (Butts, 2008).

### 3.2. 유전자-질병간 연관분석

유전자-질병간의 연관성을 분석하기 위해서 질병여부를 나타내는 변수(Group)를 포함하여 각 유전자간, 유전자 질병 간 비모수상관계수를 이용하여 네트워크 분석을 한 결과는 Figure 3.1과 같다. 유전자

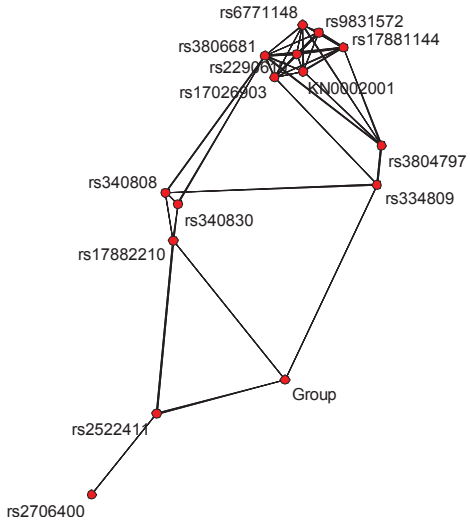



**Figure 3.1.** Network for analyzing gene-disease association

**Table 3.1.** Result of the trend test for gene-disease association

| SNP | $Z$ | $p$-value |
|---|---|---|
| rs2522411 | 3.0531 | 0.0023 |
| rs2706400 | 1.2316 | 0.2181 |
| rs17026903 | 0.2622 | 0.7932 |
| rs3806681 | −0.2867 | 0.7743 |
| KN0002001 | 0.5512 | 0.5815 |
| rs17881144 | 1.4041 | 0.1603 |
| rs6771148 | −0.1188 | 0.9054 |
| rs9831572 | 0.6479 | 0.517 |
| rs2290610 | −0.3935 | 0.694 |
| rs334809 | −1.8676 | 0.0618 |
| rs3804797 | −0.7689 | 0.4419 |
| rs17882210 | −2.1574 | 0.031 |
| rs340808 | −0.3711 | 0.7105 |
| rs340830 | −0.6823 | 0.495 |

집단은 크게 4개군 (1군-rs252241, rs2706400; 2군-rs340808, rs17882210, rs340830; 3군-rs3804797, rs334809; 4군-기타)으로 분류됨을 볼 수 있다. 질병변수와 직접적으로 연결된 유전자는 rs252241, rs17882210, rs334809의 3개 SNP이며, 위쪽의 기타군과는 직접적인 연관이 없는 것으로 보인다. rs2706400은 rs2522411을 통해서만 다른 SNP 및 질병변수로 연결되는 고립점임을 알 수 있다. 결과비교를 위하여 Armitage의 추세검정을 이용하여 질병-유전자간 연관분석을 한 결과 rs252241과 rs17882210이 유의한 것으로 나타났다($p = 0.002$, $p = 0.031$). 한편 질병변수와 직접 연결되어 있던 SNP 중 rs334809는 $p = 0.062$로 나타났다 (Table 3.1).

Table 3.2는 Figure 3.1의 네트워크에 대한 중심성 분석 결과이다. 근접 중심성은 다른 모든 개체들간

**Table 3.2.** Result of centrality analysis(top 5 SNPs)

| rank | closeness centrality | | betweenness centrality | |
|---|---|---|---|---|
| 1 | rs334809 | 6.63 | rs334809 | 87 |
| 2 | rs3806681 | 6.15 | Group | 65 |
| 3 | rs340808 | 6.00 | rs3806681 | 63 |
| 4 | Group | 5.94 | rs17026903 | 44 |
| 5 | rs17026903 | 5.91 | rs340808 | 43 |

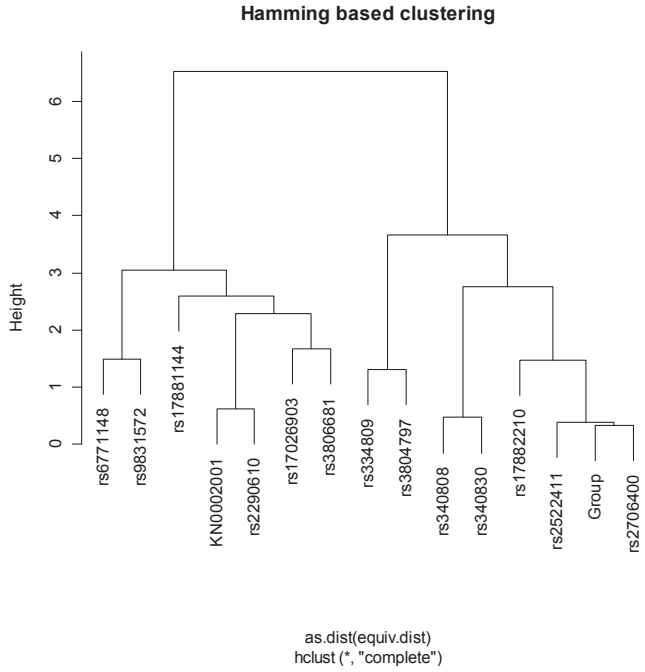



**Figure 3.2.** Result of clustering for structural equivalence

의 거리가 최단거리인, 전체 네트워크 상 거리의 중심에 위치하는 개체를 찾아낼 수 있다. 중개 중심성 지수는 다른 개체들의 최단 거리 사이에 위치하는 경우의 수로서, 특정 개체가 네트워크 내에서 중개(hub)의 역할을 많이 하느냐를 파악할 수 있다. 두 지수 공히 rs334809가 중심역할을 하고 있음을 나타낸다. 서로 직접적인 연관이 없는 SNP들이 중개 SNP을 통해서 연결되어 있을 수 있다. 따라서 중개 SNP은 치료 및 진단시에 표적이 될 수 있으며, 이를 표적하여 중재(targeted intervention)함으로써 큰 치료효과를 볼 수 있다.

Figure 3.2는 Hamming 거리를 이용하여 구조적 동치성을 기준으로 형성한 군집분석의 결과이다. 군집분석은 최장연결법에 의한 계보적 군집방법을 이용하였다. 질병변수와 rs2706400, rs2522411이 한 그룹으로 묶였다. 또한 rs340808과 rs340830, KN002011과 rs2290610이 각각 가까이 묶여서 네트워크내의 역할이 유사하여 상호대체 가능한 SNP임을 알 수 있다. 구조적으로 동치인 SNP들을 파악함으로써 직접적인 상관이 없더라도 유사한 역할을 하는 SNP군을 파악할 수 있으며, SNP의 수가 많을 때 이들의 그룹으로 다시 네트워크를 표현할 수 있게 된다.

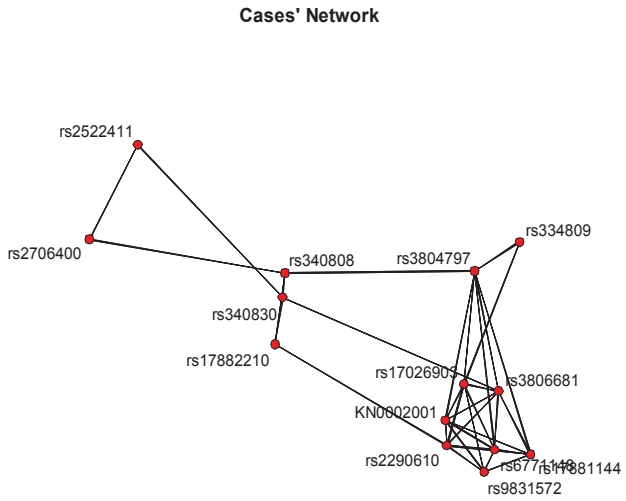



**Figure 3.3.** Gene network(case)

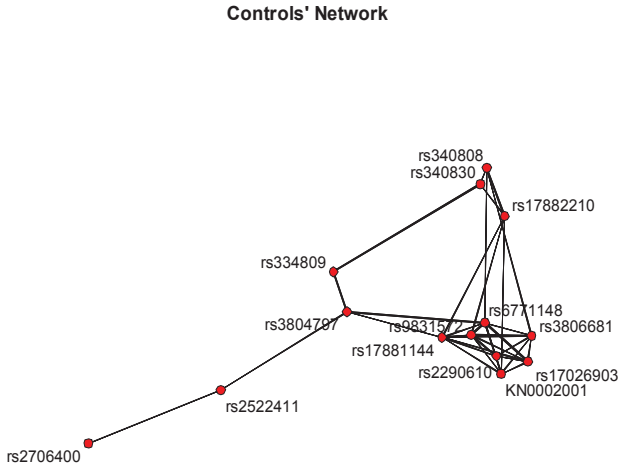


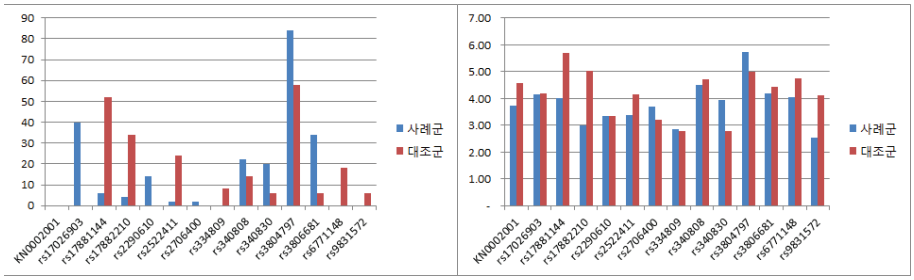

**Figure 3.4.** Gene network(control)

### 3.3. 유전자-유전자간 상호작용분석

자료를 사례군과 대조군으로 나누어 각 그룹에서 비모수상관계수 행렬을 각각 구하고 네트워크 분석을 수행하여 분석 결과를 비교하였다. Figure 3.3과 Figure 3.4는 각각 사례군과 대조군에서의 유전자간 네트워크를 보여준다. 전체 네트워크를 생성했을 때와 마찬가지로 각 그룹에서의 유전자그룹은 대략 4개의 그룹으로 묶여진다. 크게 달라지는 점은 rs2522411(1군)과 rs340830(2군)간의 관계가 대조군에서는 직접 연결되지 않고 rs3804797(3군)을 통해서 이루어지는데 반해, 사례군에서는 직접 연결되는 것을 알 수 있다. 따라서 관계에 변화를 일으킨 1군과 2군, 그리고 1군과 3군의 SNP들 간의 상호작용이

**Table 3.3.** Result of gene-gene interaction analysis using MDR

| dimension | selected model | CVC |
|---|---|---|
| 1-way | (rs2522411) | 7.0 |
| 2-way | (rs2522411, rs340808) | 5.9 |
| 3-way | (rs2290610, rs17882210, rs340808) | 3.5 |




(a) (b)

**Figure 3.5.** Centrality measure: (a) closeness centrality, (b) betweenness centrality

질병과 연관이 있음을 짐작할 수 있다. 또한 rs2706400의 경우 대조군에서는 rs2522411과만 연결이 되어 있으나 사례군에서는 rs2522411외에 rs340808과도 연결되었음을 알 수 있다.

Table 3.3은 MDR을 이용하여 유전자-유전자간 상호작용을 분석한 결과이다. 2차 상호작용의 경우 (rs2522411, rs340808)이 최적의 쌍으로 선정되었다. 3차 상호작용은 (rs2290610, rs17882210, rs340808) 조합으로 선정되었으나, CVC(Cross-validation consistency)는 3.5로 높지 않다. CVC는 MDR 실행시 데이터를 10등분하여 교차평가할 때 10회 중 최적의 모형으로 뽑히는 횟수를 의미하며, 여기서는 10회 반복시의 평균을 제시한 것이다 (Namkung 등, 2009). 1차원 상호작용, 즉 주효과의 경우 rs2522411이 선택되었는데 이는 네트워크분석에서 질병과 가장 관련이 있던 SNP으로 나타났던 rs252241, rs17882210, rs334809 중 하나이다. 또한 2차원에서의 최적 조합인 (rs522411, rs340808) 쌍의 경우 네트워크분석시 보이는 1군과 2군의 조합중 하나임을 알 수 있다. 3차원 상호작용의 경우 두 분석 모두에서 뚜렷한 해석을 할 수 없다고 판단된다. MDR의 경우 통상 최적의 조합만을 구하게 되는데, 실제로는 최적이 아니더라도 비슷한 연관성 및 상호작용을 보이는 조합들이 있는 경우가 많다. 네트워크분석에서는 이러한 관계를 하나의 그래프안에서 확인할 수 있는 장점이 있다.

Figure 3.5는 사례군과 대조군에서의 중심성지수를 보여준다. 근접중심성의 경우 대조군에서는 rs17881144 가장 높은 순위를 나타냈지만 사례군에서는 상대적으로 낮아진 값을 갖는다. 대신 rs3804797이 높은 순위를 차지하였다. 따라서 근접중심성 관점에서 사례군과 대조군의 두 네트워크의 중심이 변화했다는 것을 알 수 있다.

중개중심성의 경우 두 군 공히 rs3804797이 가장 높으며 대조군에서 높은 값을 보이는 rs17881144와 rs17882210이 사례군에서는 낮아졌다. rs17026903과 rs3806681은 사례군에서 높은 중개중심성 값을 갖는다. 이들 결과를 종합해 볼 때 중개중심성의 관점에서 rs3804797은 양쪽 모두에서 중심역할을 하고, rs17026903과 rs3806681이 사례군의 중심에, rs17882210과 rs17881144이 대조군의 중심에 특징적으로 위치한다고 해석할 수 있다.

각 네트워크에서 각 유전자의 역할이 얼마나 달라졌는지를 파악하기 위해 각 유전자 별로 두 가지 네트워크에서의 중심지수값의 차이를 구하였다. 중심점 지수에서 변화의 폭이 크다면 해당 유전자를 중

**Table 3.4.** Difference of centrality measure between case and control(top 5 SNPs)

| rank | closeness | | betweenness | |
|---|---|---|---|---|
| | SNP | difference | SNP | difference |
| 1 | rs334809 | 6.63 | rs334809 | 87 |
| 2 | rs3806681 | 6.15 | Group | 65 |
| 3 | rs340808 | 6.00 | rs3806681 | 63 |
| 4 | Group | 5.94 | rs17026903 | 44 |
| 5 | rs17026903 | 5.91 | rs340808 | 43 |
| mean($n = 15$) | −0.40 | | 0.14 | |
| standard deviation($n = 15$) | 0.93 | | 23.92 | |

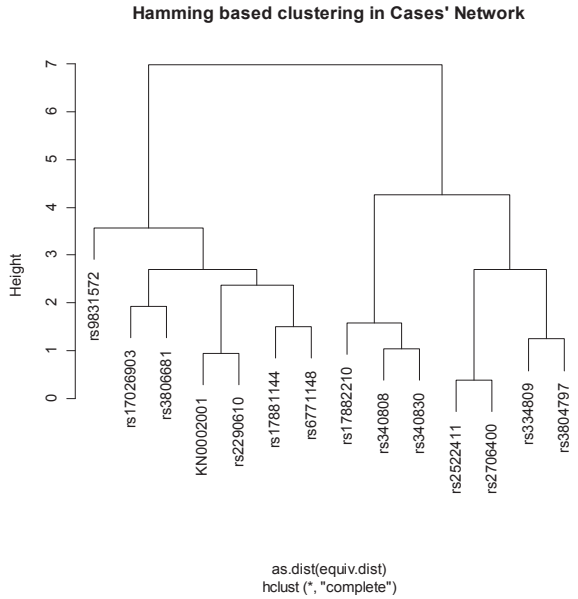



**Figure 3.6.** Result of clustering for structural equivalence(case)

심으로 하는 유전자간 상호작용이 질병과 관련이 있을 것으로 추측할 수 있다. Table 3.4는 두 네트워크에서 각 유전자 별 근접중심성과 중개중심성 지수 변화 폭이 큰 SNP들을 보여준다. rs17882210과 rs17881144의 변화가 크게 나타났음을 알 수 있다. 중심성 지수 변화량의 통계적 유의성을 알아보기 위해 종종 중심지수에 대한 짝지은 $t$-검정 등이 수행되기도 한다. 그러나 지수변화량의 평균이 0과 유의한 차이가 없다는 것이 지수변화가 없다는 의미가 되지 못하므로 이러한 분석은 적절하지 않다. 이보다는 이들의 차이에 대한 표준편차를 기술하는 것이 보다 적절하다. Table 3.4에 15개 SNP의 지수변화량의 평균과 표준편차가 있다. 근접중심성의 표준편차는 크지 않은 반면, 중개중심성의 표준편차는 커서 중개중심의 지수변화가 크다는 것을 알 수 있다.

Figure 3.6과 Figure 3.7은 대조군과 사례군에서 해밍거리에 의해 군집분석을 수행한 결과이다. 대부분의 유전자들의 구조적 동일성은 변화하지 않았다. rs9831572의 경우는 대조군에서는 rs6771148과 그 역할이 비슷했던 반면, 사례군에서는 유사한 역할을 하는 유전자를 찾기 어려웠다.

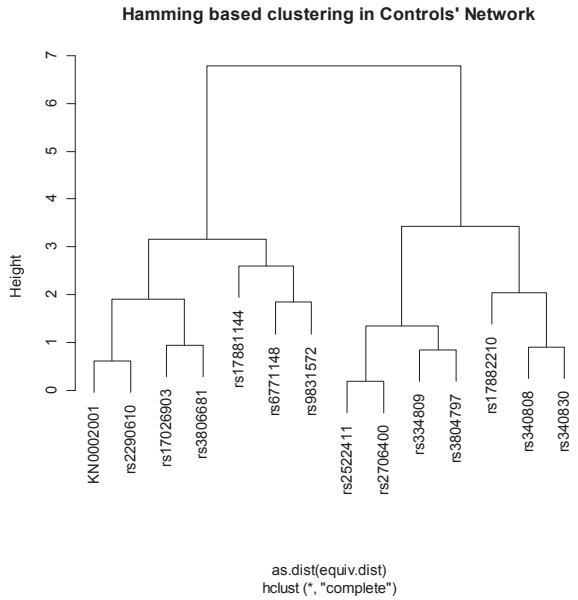



**Figure 3.7.** Result of clustering for structural equivalence(control)

## 4. 토의

본 연구에서는 사회적 네트워크분석을 이용하여 유전적 연관성 및 유전자-유전자간 상호작용을 탐색하는 방법을 제시하고, 실제 유전자데이터에 적용한 결과를 제시하였다. 네트워크 분석은 두 부분으로 나누어서, 질병여부를 함께 분석한 네트워크를 통해서는 질병에 독자적인 영향(주효과)을 미치는 SNP을 탐색하였다. 사례군과 대조군을 별도로 분석한 네트워크의 비교를 통해서는 두 네트워크에서 관계가 상이한 SNP조합을 탐색함으로써 유전자-유전자간 상호작용을 탐색할 수 있었다. 또한 각 상황에서 영향력이 있고 중심이 되는 유전자와 그들이 구성하는 집단간의 특성을 탐색할 수 있었다. 질병-유전자간 연관성을 분석하는 방법으로 가장 흔히 사용되는 추세검정결과와 비교했을 때, 추세검정결과 유의하게 나타난 SNP이 네트워크상에서 질병그룹과 직접 연결된 것으로 나타났으며, 5% 유의수준에서 유의하지 않은 것이라 하더라도 어느 정도 연관이 있는 SNP들도 동시에 탐색할 수 있어서 이러한 접근방식이 의미 있음을 알 수 있었다. 유전자-유전자간 상호작용분석에서 많이 사용되고 있는 MDR결과와 비교하였을 때, MDR에서 최적으로 뽑은 SNP조합들을 그래프 상에서도 확인할 수 있었으며, 최적이 아닌 조합이더라도 관계의 정도를 파악할 수 있었다. MDR의 경우 exhaustive한 방법으로서 최적의 조합을 찾기 위해서는 모든 가능한 쌍에 대해 10-fold 방식으로 접근하게 되므로 수많은 계산을 필요로 한다. 반면 네트워크방식은 1회의 분석으로 완료되어 시간과 노력이 훨씬 줄어든다. 또한 정보를 그래프로 표현함으로써 보다 직관적으로 데이터탐색이 가능하다는 장점이 있다. 네트워크방법의 단점으로는 이것이 탐색적인 접근방식으로서 통계적 유의성을 구체적으로 찾기는 어렵다는 것이다.

최근 Davis 등 (2010)은 사회네트워크분석의 개념을 접목하여 Marlon Brando같이 표현형에 강한 개인적 역할을 하는 SNP(Brando SNP)과 Kevin Bacon같이 다른 배우와의 연결을 통해 능력을 인정받는 SNP(Bacon SNP)을 찾는 것을 목표로 하는 GAIN(genetic association interaction network)을 제안하였다. 이 네트워크는 주효과의 강도를 노드의 가중치로, 유전자간 상호작용의 강도를 edge의 가중치로

하여 작성된다. 본 연구에서 사용된 방법이 구현된 네트워크를 통해 주효과와 상호작용을 탐색하는 것인데 반해, 이 방법은 주효과와 상호작용을 사전에 정의하여 값을 구하고 이를 그래프로 표현하는 방법론이라고 할 수 있다.

여기서는 네트워크 데이터를 생성할 때 유전자형에서의 일치 및 비일치하는 대립형질쌍의 수를 이용한 Kendall의 tau-b 값을 썼으나, IBS(identical by state)점수 등의 다양한 거리척도를 사용할 수도 있을 것이다. 네트워크를 통한 탐색적 접근을 통해 주목할 만한 SNP 및 집단들을 선별하고 추후 이들을 중심으로 하는 정밀한 실험 및 분석에 들어갈 수 있을 것이다.

## References


Balding, D. J. (2006). A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, **7**, 781–791.

Butts, C. T. (2008). Social network analysis with sna, *Journal of Statistical Software*, **24**.

Cordell, H. J. (2009). Detecting gene-gene interaction that underlies human diseases, *Nature Reviews Genetics*, **10**, 392–403.

Davis, N., Crowe, J., Pajewski, N. and McKinney, B. (2010). Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine, *Genes and Immunity*, **11**, 630–636.

Fortes, M., Reverter, A., Zhang, Y., Collis, E., Nagaraj, S. H., Jonsson, N., Prayaga, K., Barris, W. and Hawken, R. (2010). Association weight matrix for the genetic dissection of puberty in beef cattle, *PNAS*, **107**, 13642–13647.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement, *Software-Practice and Experience*, **21**, 1129–1164.

Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods, Software-Practice and Experience.

Huh, M. (2010). *Introduction to Social Network Analysis Using R*, Free Academy, Kyunggi.

Jung, J., Yee, J., Lee, S. and Park, M. (2011). Exploration of the gene-gene interactions using the relative risks in distinct genotypes, *The Korean Journal of Applied Statistics*, **24**, 861–869.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs, *Information Processing Letters*, **31**, 7–15.

Lee, H., Kim, M. and Park, M. (2010). A review of genetic association analyses in population and family based data: Methods and software, *The Korean Journal of Applied Statistics*, **23**, 95–111.

Lindlof, A. and Olsson, B. (2002). Could correlation-based methods be used to derive genetic association networks?, *Information Sciences*, **146**, 103–113.

Namkung, J. H., Kim, K., Yi, S., Chung, W., Kwon, M. S. and Park, T. (2009). New evaluation measures for multifactor dimensionality reduction classifiers in gene-gene interaction analysis, *Bioinformatics*, **25**, 338–345.

Namkung, J. H., Lee, J., Kim, E., Cho, H. J., Kim, S., Shin, E. S., Cho, E. Y. and Yang, J. M. (2007). IL-5 and IL-5 receptor alpha polymorphisms are associated with atopic dermatitis in Koreans, *Allergy*, **62**, 934–942.

Ritchie, M. D., Hahn, L., Roodi, L., Bailey, L., Dupont, W., Parl, F. and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *American Journal of Human Genetics*, **69**, 138–147.

Snel, B. and Huynen, M. (2002). The identification of functional modules from the genomic association of genes, *PNAS*, **99**, 5890–5895.

Sohn, D. (2002). *Social Network Analysis*, Kyungmunsa, Seoul.

Yanai, I. adn DeLisi, C. (2002). The society of genes: networks of functional links between genes from comparative genomics, *Genome Biology*, **3**, research0064.1-0064.12.

Zhu, X., Gerstein, M. and Snyder, M. (2007). Getting connected: analysis and principles of biological networks, *Genes and Development*, **21**, 1010–1024.