DOI: 10.3745/KIPSTD.2011.18D.2.089

단백질 상호작용 네트워크를 통한 유전체 단위반복변이와 트랜스유전자 발현과의 연관성 분석

박 치 현[†] · 안 재 균[†] · 윤 영 미[††] · 박 상 현[†††]

요 약

인간 유전체에 존재하는 유전적 구조 변이(genetic structural variation) 중 하나인 유전체 단위반복변이(Copy Number Variation, CNV)은 유전자의 기능 발현과 밀접한 관련이 있다. 특히 특정 유전 질병이 있는 사람들을 대상으로 CNV와 유전자발현의 관계를 밝히는 연구가 계속 진행되고 있지만, 정상인 유전체에 대한 CNV의 기능적 분석은 아직 활발히 이루어지고 있지 않다. 본 논문에서는 다수의 정상인 샘플에서 찾아낸 공통된 CNV에 대하여 유전자들과의 기능적 관계를 유전자의 분자적 위치와 상관없이 밝힐 수 있는 분석 방법을 제시한다. 이를 위해 서로 다른 이질적인 생물학데이터를 통합하는 방법을 제시하고 공통된 CNV와 유전자와의 연관성을 분자적 위치와 상관없이 계산할 수 있는 새로운 방법을 제시한다. 제안된 방법의 유의성을 보이기 위해서 유전자 온톨로지 (Gene Ontology) 데이터베이스를 이용한 다양한 검증 실험들을 수행하였다. 실험결과 새롭게 제안된 연관성 측정방법은 유의성이 있으며 공통된 CNV와 강한 연관성을 갖는 유전적 기능의 후보들을 시스템적으로 제시할 수 있는 것으로 나타났다.

키워드 : 유전체 단위반복변이, 트랜스유전자, 전장유전체연관분석연구, 단백질상호작용네트워크

Genome-Wide Association Study between Copy Number Variation and Trans-Gene Expression by Protein-Protein Interaction-Network

Chihyun Park[†] · Jaegyoon Ahn[†] · Youngmi Yoon[††] · Sanghyun Park[†††]

ABSTRACT

The CNV (Copy Number Variation) which is one of the genetic structural variations in human genome is closely related with the function of gene. In particular, the genome-wide association studies for genetic diseased persons have been researched. However, there have been few studies which infer the genetic function of CNV with normal human. In this paper, we propose the analysis method to reveal the functional relationship between common CNV and genes without considering their genomic loci. To achieve that, we propose the data integration method for heterogeneity biological data and novel measurement which can calculate the correlation between common CNV and genes. To verify the significance of proposed method, we has experimented several verification tests with GO database. The result showed that the novel measurement had enough significance compared with random test and the proposed method could systematically produce the candidates of genetic function which have strong correlation with common CNV.

Keywords : Copy Number Variation, Trans-gene, Genome-Wide Association Study, Protein-Protein Interaction Network

1. 서 론

유전체 단위반복변이는 가장 대표적인 유전체 구조적인 변이 중 하나로, 최근 유전체학 연구 분야에서 가장 많은 관심의 대상이 되고 있는 연구 분야이다. 2003년 종료된 휴먼 게놈 프로젝트[1]를 통해서 인간의 전체 DNA 염기서열이 결정되었고, 이후 유전체 내에 존재하는 유전자, 유전 조절 부위, 구조적 변이 등과 같은 모듈들에 대한 분석을 통하여 그 기능을 밝히려는 연구의 필요성이 증가하고 있다. 대표적으로 서로 다른 인간 유전체 내에 존재하는 다형성(Polymorphism)의 하나인 단일염기변이(Single Nucleotide Polymorphism, SNP)를 통하여 병리적 혹은 형질적 특징을 분석하는 연구가 많이 진행되어 왔다[2][3]. 기존의 연구들은 찾아낸 SNP의 존재를 통하여 특정 질병을 갖는 그룹에 대한 유전적 표식(Genetic Marker)으로 활용해왔다. 실제로 다

※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0008639).
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2010-0010689).
† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
†† 종신회원 : 가천의과학대학교 정보공학과 교수
††† 종신회원 : 연세대학교 컴퓨터과학과 교수(교신저자)
논문접수 : 2010년 11월 26일
수 정 일 : 1차 2011년 2월 14일, 2차 2011년 2월 15일
심사완료 : 2011년 2월 16일

수의 논문에서 SNP의 영향으로 주변 유전자(Gene)의 발현이 조절되는 경우를 실험을 통하여 밝히고 있다.

하지만 A혹은 B의 이분법으로 나타나는 SNP을 통하여 다양한 유전적 형질 혹은 복합 질환의 메커니즘을 통합적으로 설명하는 것은 한계가 있었다. 또한 유전체 분석 기술이 발달하면서, SNP 뿐만 아니라 다른 종류의 유전체 다형성을 찾는 연구가 활발히 진행되면서 유전체 단위반복변이(Copy Number Variation 이하, CNV)의 존재가 중요시되고 있다[4][5][6]. [7]에서는 CNV란 정상 사람의 유전체(Reference sample)에 대비하여 1Kb(Kilo-basepair)이상 염기서열이 중복(gain) 혹은 손실(loss)된 상태를 의미한다고 정의했다. CNV는 일정 길이 이상으로 염기서열이 중복 혹은 손실되었기 때문에 이런 하나 이상의 유전자를 포함할 수도 혹은 발현 조절부위(Transcription Factor)를 포함할 수 도 있다[8]. 이 때문에 유전자 발현이 억제 혹은 활성화 될 가능성이 존재하며 결과적으로 CNV의 존재는 복합 유전자 질환 혹은 다양한 유전적 형질 차이의 원인이 될 수 있다[22]. 또한 CNV는 SNP과 마찬가지로 자연선택압력(Natural Selective Pressure)에 영향을 받기 때문에 그들의 기능적 작용에 의해서 서로 다른 빈도수를 나타내게 되는데 이는 통계적으로 유전질병과의 연관성을 밝히기 위해 필요한 중요한 성질이다. 예를 들어 통계적으로 A라는 질병이 B라는 CNV와 관련이 있다는 것을 밝히기 위해서는, B라는 CNV가 A라는 질병이 있는 집단에서 높은 빈도수로 존재하지만 정상인 집단에서는 낮은 빈도수로 존재해야 한다. 즉 CNV라는 것이 자연선택압력에 의해 영향을 받기 때문에 A라는 질병을 가진 집단에서는 유전적으로 B라는 CNV가 계속 존재하도록 선택되어야 한다라는 가정이 필요한데 SNP과 마찬가지로 CNV에서도 이러한 가정이 실제로 존재함이 밝혀졌다[9]. 따라서 지금까지 CNV의 기능 추론에 대한 연구는 암 혹은 복합유전질병에 관계하여 주로 수행되어 왔다[10][11][20]. 이는 위에 언급했듯이 질병이라는 특징(trait)이 있는 경우 통계적으로 질병과 연관된 집단과 CNV의 연관성을 추론하는 방법이 용이해지며, 이때의 CNV의 기능에 대해서 해당 질병과 관련된 유전자들의 발현 정도를 분석함으로써 생물학적인 검증 또한 쉽게 이루어질 수 있기 때문이다[12][20].

그러나 이러한 연구는 특정 질병과 관련된 CNV의 기능을 찾는 것이 목적이기 때문에 인간이 가지고 있는 전체 CNV를 대상으로 하지 않으며 유전체내 특정 위치에 존재하는 CNV에 대해서만 분석을 수행하는 특징이 있다. 이러한 연구 방식은 질병 각각에 대한 메커니즘은 분석할 수 있지만 아직까지 많은 복합질병은 그 발병원인이 매우 복잡하여서 이러한 단순한 분석의 합으로는 전체단위에서의 인간의 질병 메커니즘을 분석할 수는 없는 한계가 있다[6, 13]. 또한 질병에 관련된 CNV의 기능을 밝히려는 연구는 많이 수행되고 있지만, 질병이 없는 정상인 유전체에 대한 분석 연구는 아직 많이 이루어지지 못하고 있다. 정상인 유전체의 분석이 필요한 이유는 질병을 갖는 집단과의 비교 연구를 위해서는 정확한 대조군(Control)이 필요하기 때문이다. GWAS 연구의 대표적인 방법인 후향연구(Case-Control study)의 경우 희귀질병연구에 있어서 장점을 갖는 연구법인데 이런 연구법을 위해는 정상집단에 존재하는 CNV를 밝히고 그것들의 기능을 론하는 연구가 선행되어야 한다. 또한 2003년 최초 인간유체 염기서열이 밝혀진 후 시퀀스분석 기술 및 마이크로어이(microarray) 기반 기술들이 급속히 발전하여서 현 1000Genome Project, HapMap 3-phase 등 대규모 유전체 석 프로젝트들이 진행되고 있다. 이 프로젝트들은 특정 질을 가진 사람이 아닌 일반인의 유전체를 대상으로 실험을 행하고 있으며 이 과정에서 생산되는 방대한 양의 염기서정보 및 CNV를 탐지할 수 있는 aCGH(array Comparati Genome Hybridization) 데이터가 계속해서 공개되고 있 하지만 현재까지 대규모 정상인 집단에 대해서 CNV와 유자발현과의 연관성 분석을 수행한 연구는 사실상 [14]이 일하다. 최근 정상인 유전체 분석을 위한 고해상도 aCGH 이터가 출시되고 있고 최근 1Kb 이하의 작은 CNV들이 이크로어레이 기술과 시퀀싱 기술의 발달로 밝혀지고 있지 아직까지 연관성 분석을 위해 사용 가능한 다수의 샘플에 한 고해상도 aCGH 데이터는 많이 부족한 실정이다[15]. 한 이런 새로운 고해상도 aCGH 데이터에 대해서 현존하 CNV 탐색 알고리즘 또한 오차율이 높기 때문에 정확한 관성 분석을 하기 어려운 상황이다.

하지만 가장 큰 문제점은 특정 질병이 관련되어 있지 은 정상인에 대한 분석을 할 때는 이질적인 생물학 데이 (heterogeneous biological data)를 통합하여 유전체 전체 준에서 CNV와 유전자 발현과의 연관성을 분석해야 하는 이에 필요한 방법론들이 아직 충분하지 않다는 점이다[21 특히 [14]의 가장 큰 한계점은 CNV가 영향을 미치는 유 자로 해당 CNV의 앞뒤 2Mb(Mega-basepair) 이내에 존 하는 것들만을 대상으로 분석을 수행하였다. 즉 cis-actin 유전자(이하, cis-gene)에 대한 분석만을 수행하였지만, 실 로 CNV가 영향을 미치는 유전자들은 trans-acting(이하 trans-gene) 유전자들도 상당수 있을 것으로 예상되며 논문에서 밝히고 있다[10, 21]. 여기서 cis-acting이란 유 자의 활성에 영향을 미치기 위해서 같은 염색체 상에 존 해야만 하는 enhancer 또는 promoter와 같은 유전자 요소 을 일컫는 말이다. 따라서 cis-gene이란 같은 염색체 내에 이러한 유전자 요소에 의해서 발현이 조절되는 유전자를 미한다. 또한 trans-acting이란 전사인자(transcription factor 와 같이 별개의 염색체상에 존재하더라도 여전히 다른 유 자에 영향을 미칠 수 있는 유전적 인자들을 설명하는 개 이다. 따라서 trans-gene이란 다른 염색체에 존재하면서 이러한 유전적 인자에 의해 발현이 조절되는 유전자를 의미한다. 간단히 말해, 본 논문에서 쓰이는 cis-gene은 같은 염색체 내에서 CNV의 근거리에 위치한 유전자 요소들에 의해 영향을 받을 수 있는 유전자를 뜻한다. 반대로 trans-gene은 CNV의 근거리에 위치한 유전자 요소들에 의해 직접적으로 발현이 조절되지 않는 유전자를 의미한다. 이와 관련해서 [10] 논문은 cis-유전자뿐만 아니라 trans-유전자들과 CNV의 관계를 밝히는 방법론을 제시하였는데, 다만 사용한

데이터의 크기가 작은 암 관련 데이터들이었다. 따라서 현재 단계에서 CNV와 유전자발현과의 연구에 있어서 정상인과 같은 대용량 데이터에 대해서 cis-유전자들뿐만 아니라 trans-유전자과의 관계도 유추할 수 있는 방법이 필요하다. 이 과정에서 서로 특성이 다른 대용량의 이질적인 생물데이터를 효과적으로 통합하며 그 안에서 새로운 사실을 찾아낼 수 있는 분석 방법론이 필요하다.

본 논문에서는 다수의 정상인들을 대한 분석 방법을 제시하기 때문에 CNV가 아닌 CNVR(CNV Region)과 유전자들과의 기능적 관계를 추론할 수 있는 데이터 통합 방법론 및 새로운 상관관계 측정 방법을 제안한다. 제안하는 방법의 가장 큰 특징은 cis-유전자뿐만 아니라 trans-유전자들과 CNVR과의 연관관계를 단백질 상호작용 네트워크를 이용하여 추론하는 것과, 이 과정에서 사용되는 상관관계 측정 방법을 본 논문에서 사용하는 데이터의 특성에 맞추어서 새롭게 제안하는 것이다. 또한 제안하는 방법을 통해 도출된 유전자집합에 대해서 GO 데이터베이스를 이용한 기능적 연관성 검증을 수행하여 도출된 CNVR과 유전자집합 사이의 유의성을 검증한다.

2. CNVR과 유전자발현과의 연관성 분석

본 장에서는 제안하는 CNVR과 유전자집합과의 기능적인 연관관계를 분석하는 방법에 대해서 기술한다. 제 2.1 절에서는 본 연구에서 사용하는 데이터 집합에 대하여 기술한다. 제 2.2 절에서는 본 논문에서 언급하는 기본 개념에 대한 정의를 한다. 제 2.3 절에서는 전체적인 통합 분석 알고리즘에 대해서 기술한다.

2.1 사용하는 데이터 집합

본 논문에서는 크게 3가지 데이터를 사용한다. CNVR과 유전자집합과의 연관성을 밝혀야 하므로 CNVR을 찾을 수 있는 aCGH 데이터, 유전자의 발현값을 측정한 GE(Gene Expression)데이터, 마지막으로 trans-유전자 간의 관계를 유추할 수 있도록 PPI(Protein-Protein Interaction) 데이터가 사용된다. CNVR와 유전자발현과의 관계를 밝히기 위해서는 동일인에 대해서 aCGH실험과 유전자발현 실험이 수행된 데이터쌍이 있어야 하는데 정상인에 대한 이러한 데이터는 아직까지 많이 부족한 편이다. 암과 같은 질병과 관련된 데이터는 이러한 쌍 데이터가 다수 존재하지만, 본 논문에서는 정상인을 대상으로 수행하는 것이기 때문에 2006년 WTSI(Wellcome Trust Sanger Institute)에서 공개한 27000 클론으로 구성된 WGTP(Whole Genome TilePath) HapMap 샘플 210명의 aCGH 데이터를 사용하며 [16], 동일인에 대해서 측정된 GE(the transcription of DNA into mRNA)을 사용하였다. 사용된 aCGH 데이터는 [23]에서 다운로드 받을 수 있으며, GE 데이터는 Gene Expression Omnibus(GEO: submission number series: SGNE6536)에서 다운 받을 수 있다. 이 두 데이터는 [14]에서 사용된 것으로 이 이후에도 정상인에 대한 새로운 고해상도 aCGH 데이터가 공개되었지만 해당 샘플의 수가 40개로 적으며, 새로운 고해상도 데이터에 대한 GE데이터가 아직 공개되지 않았기 때문에 본 논문에서 사용할 수는 없었다. aCGH 데이터와 GE 데이터의 구조는 다음 (그림 1)과 같다.

마지막으로 PPI데이터의 경우는 Interologous Interaction Database [24]에서 다운받을 수 있는데, 효모, 쥐, 인간 등 여러 생물 종에 대한 단백질 상호작용을 측정하여 데이터베

Positions of genome

Array-CGH Data (Position)	Sample 1	sample 2	...	sample 210
Chromosome1 Probe 1 (Position: 1)	$r_{1,1}$	$r_{1,2}$	...	$r_{1,210}$
Chromosome1 Probe 2 (Position: 2)	$r_{2,1}$	$r_{2,2}$	...	$r_{2,210}$
...	...	...	...	...
Chromosome22 Probe 382 (Position: 24860)	$r_{24860,1}$	$r_{24860,2}$	...	$r_{24860,210}$
Chromosome22 Probe 383 (Position: 24861)	$r_{24861,1}$	$r_{24861,2}$	...	$r_{24861,210}$

Genes

Gene Expression (Namespace: Illumina_v1)	sample 1	sample 2	...	sample 210
GI_10047089-S	$e_{1,1}$	$e_{1,2}$	...	$e_{1,N}$
GI_10047091-S	$e_{2,1}$	$e_{2,2}$	...	$e_{2,N}$
...	...	...	...	...
thrB	$e_{47293,1}$	$e_{47293,2}$	...	$e_{47293,N}$
trpF	$e_{47294,1}$	$e_{47294,2}$	...	$e_{47294,N}$

CNVR

***vcnvr* (*22,24860,24861*)**	μ_1 = mean ($r_{24860,1}$, ~$r_{24861,1}$)	μ_2 = mean ($r_{24860,2}$, ~$r_{24861,2}$)	...	μ_{210} = mean ($r_{24860,210}$, ~$r_{24861,210}$)

(그림 1) aCGH 데이터와 GE 데이터의 구조

이스를 구축하였기 때문에 많은 타 연구에서 인용되고 있다 [17]. 특히 인간 PPI 데이터베이스의 경우 정상인을 대상으로 단백질간의 상호작용 여부를 밝힌 것이기 때문에 본 논문에서 사용하기가 적합하였다. 본 논문에서 사용하는 PPI 데이터는 2개의 단백질로 이루어진 모든 가능한 조합에 대해서 상호작용이 있을 경우 방향성 없이 두 단백질이 하나의 쌍으로 존재하는 구조로 이루어져 있다. 사용하는 데이터의 경우 이러한 단백질-단백질 쌍이 약 14만개 정도 존재한다. 본 논문에서는 위 3가지 데이터를 통합하여 분석함으로써 기존연구에서 하지 못했던 새로운 CNVR과 유전자발현과의 연관성 분석을 시도한다.

2.2 상관관계 규칙 마이닝의 응용

서로 다른 항목에 대한 연관 규칙 마이닝에서 규칙들은 지지도(support)와 신뢰도(confidence)을 임계치로 하여 결정이 된다. 하지만 지지도와 신뢰도만으로 연관 규칙을 도출하는 것은 한계가 있으며 그 대안으로 상관관계(correlation)을 이용하여 두 데이터에 대한 연관 규칙을 결정하는 방법이 있다. 이때 상관관계 분석은 PCC, *lift*, χ^2 등을 계산하여 얻은 두 항목 사이의 상관도를 이용하여 연관규칙을 도출한다. 이러한 마이닝 기법을 상관관계 분석(Correlation Analysis)이라고 하는데, 본 논문에서는 이 기법을 응용하는 새로운 연관성 분석 방법을 제안한다. 기본적으로 두 항목 A와 B 사이의 연관성 규칙은 다음과 같이 표현된다.

$$A \rightarrow B\ [support,\ confidence,\ correlation]$$

본 논문에서는 항목 A를 하나의 CNVR, 항목 B를 하나의 유전자로 맵핑하여 두 항목 사이의 상관관계를 분석한다. 하지만 항목 A와 B는 서로 도메인이 다른 데이터이며 트랜잭션 데이터의 형태가 아니기 때문에 지지도와 신뢰도를 계산하기가 어렵다. 또한 사용하는 데이터가 210개의 요소로 이루어진 만큼 PCC와 같은 선형적인 상관관계 측정 방식으로는 두 항목 사이에 존재하는 연관성을 정확히 측정하기 어렵다. 이유는 요소의 수가 많고 데이터의 특성 자체가 노이즈가 많기 때문에 숨어있는 상관도를 찾지 못하기 때문이다. 또한 본 논문에서 사용하는 데이터처럼 하나의 항목 B가 여러 하위 항목들인 b_1, b_2, b_3, … , 를 가지고 있고 항목 A와 항목 b_1의 관계도 고려를 해야 한다면 PCC나 *lift* 와 같은 방법은 정확한 상관도를 보여줄 수 없다. 따라서 본 논문에서는 숨겨진 상관도를 구하는데 특화되어 있으며, 상관분석에 사용되는 하나의 데이터항목이 하위에 또 다른 데이터항목들을 계층적으로 포함하고 있을 경우 유용하게 응용될 수 있는 새로운 상관관계측정 방식을 제시한다. 제안하는 상관관계측정 방법에 대해서는 다음 장에서 자세하게 기술한다.

2.3 기본 개념 정의

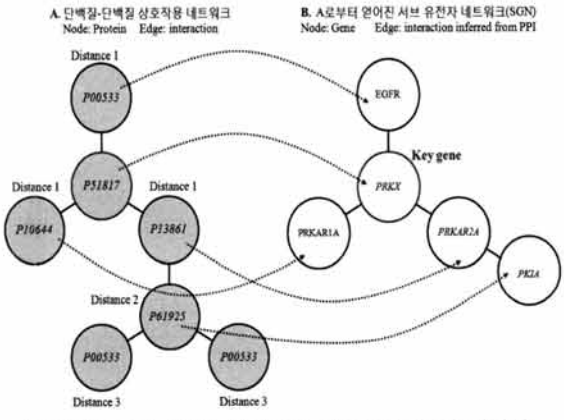
본 논문에서는 새로운 방법을 통하여 CNVR을 밝히지 않고 [7] 에서 밝힌 CNVR 결과를 그대로 이용한다. CNVR은 CNV들의 합집합을 뜻하는데, CNV란 한 사람 혹은 샘플에 대해서 정상유전체에 대비하여 1Kb(Kilo-basepair)에서 M 단위의 길이로 DNA 절편이 복사 혹은 삭제된 영역을 나타낸다. 그리고 CNVR은 각 사람 혹은 샘플에서 얻은 CNV들에 대해서 서로 1bp 라도 겹치는 부분이 있는 것들을 병합한 영역을 나타낸다. CNVR의 개념은 [7]에서 처음 제시되었으며 이후 더 정확하고 짧은 CNV 영역을 구하기 위한 여러 알고리즘들이 제시되었지만, 본 논문에서 사용한 aCGH 데이터의 경우 낮은 해상도 때문에 결과의 개선 정도가 크지 않았다. 또한 [7]에서는 알고리즘적으로 찾아낸 CNVR에 대하여 생물학적 검증을 함께 수행하였기 때문에 본 논문에서는 독자적인 CNVR 결정 알고리즘을 적용하지 않고, [7]의 CNVR 결과를 사용한다. CNVR 결과와는 별도로 전체 염색체에 대한 aCGH 발현 값들은 GE 데이터와의 연관성을 계산하는 과정에서 사용된다. CNVR외에 본 논문에서 제시하는 네 가지 새로운 개념에 대해서는 다음과 같이 정의한다.

〈표 1〉 기호와 정의

기 호	정 의
$vcnvr_{(c,\ i,\ j)} = [\mu_1, \mu_2, \cdots \mu_N]$	aCGH의 c번째 염색체의 i위치부터 j위치 사이의 log ratio들의 평균값 벡터 (N은 사용된 샘플 수)
$vg_i=[e_1, e_2, \cdots e_N]$	유전자 i에서 N개의 유전자발현 값 벡터
$SGN_{(K,\ L)}=\{g_k, g_1, g_2, \cdots\}$	유전자 K를 시작으로 PPI에서 L거리의 인접 노드(node)까지 탐색했을 경우 얻을 수 있는 서브 유전자 네트워크
PSS	벡터 X, Y의 각 대표값이 1 혹은 -1로 같은 경우 원소들의 인덱스 집합
NSS	벡터 X, Y의 각 대표값이 0이 아니면서 서로 다른 경우 원소들의 인덱스 집합
$SES(X_1, Y_1, Y_2)$	벡터 X_1와 Y_1의 PSS, NSS와 벡터 X_1와 Y_2의 PSS, NSS 사이의 최대 유사도 (Selective Element-set Similarity)
r_{ij}	염색체 상 위치 i의 샘플 j에서의 aCGH의 측정값 (log ratio of intensity)
e_{ij}	유전자 i에서 샘플 j에서의 유전자발현 측정값
a_i	i번째 샘플에서 변환된 대표값
N	전체 샘플의 수
Sim	하나의 $SGN_{(K,\ L)}$에 대해서 구할 수 있는 SES들의 평균
$t(a)$	유의수준 α에 해당하는 *Sim* 값의 임계치

[정의 1] ($SGN_{(K, L)}$). $SGN_{(K, L)}$는 단백질 네트워크에서 단백질을 유전자로 대체한 유전자 네트워크에서, 유전자 K를 중심으로 거리가 L이하인 서브 유전자 네트워크이다.

이때 단백질과 유전자 사이의 관계는 각 단백질이 발현될 경우 필요한 RNA 전사에 관여하는 유전자들을 구함으로서 알 수 있다. 우리는 PPI를 통해서 단백질 네트워크를 구할 수 있으며 이는 다시 유전자 네트워크로 치환이 가능하다. 본 논문에서 위와 같은 방법으로 $SGN_{(K, L)}$를 구하는 이유는, 서로 기능적으로 연관되어있을 가능성이 큰 유전자집단을 알아내기 위해서이다. 결과적으로 유전자의 기능이라는 것은 단백질에 의해서 발현되기 때문에 PPI를 통해서 유전자들을 특정 집합으로 묶을 수 있다. 또한 이렇게 구해진 각 $SGN_{(K, L)}$에는 cis-유전자와 trans-유전자들이 모두 포함될 수 있기 때문에 CNVR과의 연관성을 유전자의 위치에 상관하지 않고 유추할 수 있다. 본 논문에서 사용하는 단백질 네트워크는 노드(node)의 개수가 약 1만3천개, 엣지(edge)는 약 14만개로 이루어졌다. 전체 네트워크의 크기가 매우 크기 때문에 어떤 한 노드에서 인접 노드들을 탐색할 때 탐색 중단 조건을 주지 않으면 탐색되는 노드들의 범위가 커질 수 있다. 본 논문에서는 모든 유전자들을 키(key)유전자로 하여서 해당 키 유전자로부터 2-거리까지 존재하는 인접 노드들을 탐색하여 SGN를 구한다. (그림 2)와 같이 단백질 네트워크(A)로부터 유전자들과의 맵핑 관계를 이용하여 유전자 네트워크를 유추할 수 있는데 이 때 2-단계까지 탐색을 허용한 경우 SGN를 구해보면 (B)를 얻을 수 있다.



(그림 2) 거리 2까지 탐색했을 경우 PPI로부터 얻어진 $SGN_{(PRKX, 2)}$의 예

[정의 2] (PSS) 모든 샘플 N에 대해서 벡터 X의 i번째 샘플의 대표값과 벡터 Y의 i번째 샘플의 대표값이 서로 0이 아니면서 같은 값을 갖는 원소들의 인덱스 집합.

PSS는 다음과 같이 표현된다.

$$PSS = \{\mathrm{i} | (\mathrm{x}_\mathrm{i} = 1 \text{and } y_i = 1) \text{or } (x_i = -1 \text{and } y_i = -1)\}, i \le N$$

[정의 3] (NSS) 모든 샘플 N에 대해서 벡터 X의 i번째 샘플의 대표값과 벡터 Y의 i번째 샘플의 대표값에서 둘 중 하나가 0이 아니면서 서로 다른 값을 갖는 원소들의 인덱스 집합.

NSS는 다음과 같이 표현된다.

$$NSS = \{\mathrm{i} | (\mathrm{x}_\mathrm{i} = 1 \text{and } y_i = -1) \text{or } (x_i = -1 \text{and } y_i = 1)\}, i \le N$$

PSS와 NSS를 결정하기 위해 두 벡터 대해서 각각의 원본 값을 -1, 0, 혹은 1로 나타나는 대표값으로 변환한 후 PSS, NSS 집합을 구한다. 두 벡터에 대한 대표값은 K-평균 군집화(K-means clustering) 방법을 통하여 얻는다. 각 벡터의 값에 대한 K=3으로 K-평균 군집화를 수행하면 각 벡터의 요소값들은 3개의 군집으로 나뉘게 되고, 중심값의 크기에 따라서 -1, 0, 1로 대표값을 결정한 후 각 군집에 속하는 요소값들을 대표값으로 변환시킨다. PSS와 NSS는 대표값이 1 혹은 -1일 경우에만 고려하여 모든 샘플들에 대해서 X의 i번째 샘플의 대표값과 Y의 i번째 샘플의 대표값이 서로 값이 같으면 PSS로, 다르면 NSS로 그 샘플을 분류한다. PSS는 각 샘플에 대해서 1 혹은 -1로 같은 값을 가지는 샘플들이므로 이 샘플들의 분포만 찍어보면 X와 Y의 실제 값에서 업-다운 패턴이 비슷한 샘플들만 뽑아낸 것이라는 것을 알 수 있다. 반대로 NSS의 경우는 서로 다른 대표값을 가지는 것들이기 때문에 업-다운 패턴이 반대로 발생하는 샘플들만 선택적으로 뽑은 것이다. 본 논문에서는 가능한 한 전체 샘플에서 긍정연관성(Positive Correlation)을 갖는 샘플들과 부정연관성(Negative Correlation)을 갖는 샘플들만 추려내서 더욱 정확하게 X와 Y, 즉 CNVR과 유전자와의 발현의 상관관계를 조사하고자 했기 때문에 이런 방법을 적용하였다.

[정의 4] ($SES(X_1, Y_1, Y_2)$). $SES(X_1, Y_1, Y_2)$는 벡터 X_1와 Y_1의 PSS, NSS와 벡터 X_1와 Y_2의 PSS, NSS 사이에서 구할 수 있는 가능한 모든 유사도의 최대값을 나타낸다.

$SES(X_1, Y_1, Y_2)$는 본 논문에서 새롭게 정의한 유사도 측정 방법이다. $SES(X_1, Y_1, Y_2)$는 주어진 3가지 벡터로부터, X1과 Y1에서 PSS1, NSS1을 구하고 X_1과 Y_2에서 PSS_2, NSS_2를 구한다. 이렇게 구해진 2개의 PSS와 2개의 NSS에 대해서 가능한 두 집합에 대해서 교집합의 비율을 구한 후 그 값들 중 최대값을 $SES(X_1, Y_1, Y_2)$로 결정한다. 수식으로 다음과 같이 정의된다.

$$SES(X_1, Y_1, Y_2) = \max \begin{cases} \dfrac{|\mathrm{PSS}_1 \cap \mathrm{PSS}_2|}{N} \\ \dfrac{|\mathrm{PSS}_1 \cap \mathrm{NSS}_2|}{N} \\ \dfrac{|\mathrm{NSS}_1 \cap \mathrm{PSS}_2|}{N} \\ \dfrac{|\mathrm{NSS}_1 \cap \mathrm{NSS}_2|}{N} \end{cases} \quad \text{(식 1)}$$

본 논문에서 이런 새로운 유사도 측정 방법을 사용한 이유는 PCC(Pearson's Correlation Coefficient)와 같은 상관관계 계수는 참여하는 두 변수에서 참여하는 요소들 중 전체 데이터의 상관성을 해치는 것들에 민감하게 값이 변하기 때문이다. 특히 aCGH 데이터의 경우 특히 노이즈가 많기 때문에 단순히 PCC를 이용할 경우 GE와의 연관성이 잘 드러나지 않을 수 있다. 따라서 N개의 샘플에서 부분 집합인 PSS, NSS을 구해서 그 안에서 대표값들의 유사도를 측정함으로써 두 변수 사이의 숨겨진 연관성을 알아낼 수 있다.

2.4 SES를 이용한 연관성 분석 알고리즘

전체 연관성 분석 방법은 기존 연구와 다르게 trans-유전자들에 대해서도 CNVR과의 연관관계를 알아봐야 하기 때문에 모든 가능한 유전자에 대해서 분석을 수행한다. 첫째로 PPI 데이터에서 모든 유전자를 키유전자로 가정한 후 $SGN_{(K,\ 2)}$를 구한다. 본 논문에서는 주어진 모든 가능한 키유전자로부터 거리 2까지 넓이우선탐색(Breadth First Search)를 수행함으로써 각 키유전자에 대한 SGN들을 구한다. 이때 탐색 결과 인접 노드 찾을 수 없는 키유전자들은 전체 SGN 집합에서 제거한다. 둘째로 [7]에서 밝힌 CNVR과 210 샘플의 aCGH 데이터를 이용하여 $vcnvr_{(c,\ i,\ j)}$을 만든다. 셋째로 각 $vcnvr_{(c,\ i,\ j)}$에 대해서 step 1에서 구한 모든 SGN들에 대해서 $SES(X_1, Y_1, Y_2)$을 구한 후 하나의 키유전자에 대해서 구해진 모든 $SES(X_1, Y_1, Y_2)$의 평균 값이 t(α) 이하인 경우는 결과에서 제거한다. 전체 연관성 분석 과정을 알고리즘으로 기술하면 아래와 같다.

$SES(X_1, Y_1, Y_2)$의 계산은 각 $SGN_{(K,\ 2)}=\{g_k, g_1, g_2, \cdots\}$에서 키유전자, vg_k 와 $vcnvr_{(c,\ i,\ j)}$와의 PSS와 NSS를 구하는 것으로부터 시작한다. 이때 $SES(X_1, Y_1, Y_2)$의 X_1은 $vcnvr_{(c,\ i,\ j)}$이 되며 Y_1은 vg_k가 된다. CNVR과 g_k사이의 PSS, NSS가 결정되면 이제 집합 SGN의 나머지 유전자들 $g_1, g_2, \cdots$에 대해서 각 유전자를 Y_2로 간주하여 각각 $vcnvr_{(c,\ i,\ j)}$과의 PSS, NSS를 구하게 된다. 본 논문에서는 이때 얻어지는 g_1에 대해서 얻은 PSS, NSS를 각각 PSS_1과 NSS_1이라고 하며, 나머지 유전자에 대해서 얻을 수 있는 PSS, NSS를 각각 PSS_2과 NSS_2이라고 한다. $SES(X_1, Y_1, Y_2)$을 구한 후 나머지 각 유전자에 대해서 PSS, NSS를 구하면서 최종적으로 $SES(X_1, Y_1, Y_2)$값들의 평균을 구한다.

예를 들어 키유전자를 제외한 나머지 유전자로 g_1, g_2가 있다고 한다면 $SES(CNVR, vg_k, vg_1)$, $SES(CNVR, vg_k, vg_2)$를 구하게 되고, 이 두 값의 평균값이 최종적으로 $vcnvr_{(c,\ i,\ j)}$과 g_k에 대한 유사도 Sim이 된다. Sim은 아래 나오는 (식 2)와 같이 정의된다.

Input:	PPI data, aCGH data, GE data, CNVR_result, Protein_To_Gene map information, t(α)
Output:	*Sim* for each CNVR
1	**For** each protein, p_i in PPI
2	p_i is converted g_i using Protein_To_Gene map information
3	**Store** g_i to PPI data instead of p_i
4	**End For**
5	**For** each gene, g_i, in GE data
6	**Set** g_i to K and 2 to L
7	Find neighbour-genes with L using BFS approach in PPI data
8	**If** neighbour-genes is not empty **then**
9	**Create** $SGN_{(K,\ L)}$ with found neighbour-genes
10	**End For**
11	**For** each CNVR in CNVR_result
12	**Set** C to the chromosome number
13	**Set** i to the start position and j to the end position of CNVR
14	**Create** $vcnvr_{(c,\ i,\ j)}$ using aCGH data
15	**Set** X_1 to $vcnvr_{(c,\ i,\ j)}$
16	**For** each *SGN* in all $SGN_{(K,\ L)}$
17	**Set** *Sim* to 0 and *sum_Sim* to 0
18	**Set** Y_1 to g_k of *SGN*
19	**For** each neighbour-genes, g_i, in *SGN*
20	**Set** Y2 to g_i
21	Calculate $SES(X_1, Y_1, Y_2)$
22	**If** the value of $SES(X_1, Y_1, Y_2) > t(\alpha)$ **then**
23	Add the value of $SES(X_1, Y_1, Y_2)$ in *sum_Sim*
24	**End For**
25	**End For**
26	Calculate *Sim* using *sum_Sim*
27	**End For**

$$Sim\left(vcnvr, SGN_{(g_k,2)}\right) = \frac{\sum_{i=1}^{h} SES(vcnvr, vg_k, vg_i)}{h} \quad (식\ 2)$$

(식 2)에서 h는 하나의 $SGN_{(K,\ 2)}$에서 키유전자를 제외한 하위 유전자들의 개수를 나타낸다.

제안하는 방법은 모든 $SGN_{(K,\ 2)}$에 대해서만 *Sim*을 계산하기 때문에 인접 노드가 하나도 없는 유전자들은 처음부터 CNVR과의 연관성을 검증하지 않게 된다. 이유는 이러한 유전자를 키유전자로 하는 $SGN_{(K,\ 2)}$을 만들어지지 못하기 때문이다. 이런 방법을 적용한 이유는 단순히 CNVR과 하나의 유전자의 연관성을 PCC 등으로 검증하는 것은 서로 거리가 멀리 떨어진 유전자의 경우 정확한 연관성을 측정하기 어렵기 때문이다. 즉 인접노드가 없이 키유전자만 있는 $SGN_{(K,\ 2)}$를 고려한다면, 단순히 CNVR과 하나의 유전자와의 연관성만 구하는 것인데 CNVR과 GE 값들이 항상 선형적으로 관계가 있지 않다. 즉 CNVR과 GE사이에 어떤 기능적 연관관계가 있다고 하여서 항상 PCC와 같은 연관성계수가 1에 가까이 커진다든지 -1에 가깝게 값이 작아지는 것은 아니다. 따라서 본 논문에서는 모든 유전자들에 대해서 가능한 연관성을 찾되 잘못 찾는 비율(False Rate)을 줄이기 위해서 CNVR과 하나의 유전자와의 연관성 정도를 측정할 때 그 유전자와 비슷한 기능을 하는 다른 유전자들과도 한번 연관성을 검사함으로써 오류를 줄인다.

최종적으로 모든 가능한 $SGN_{(K,\ 2)}$에 대해서 구한 모든 *Sim*에 대해서 높은 연관성을 갖는 키유전자들을 선택하기 위해서 *t(a)*를 적용한다. *t(a)*는 임의로 선택된 $vcnvr_{(c,\ i,\ j)}$과 $SGN_{(K,\ 2)}$사이에서 얻어진 *Sim*에 대해서 *a* 유의수준에 해당하는 값을 의미한다. *t(a)*의 결정 방법은 실험 3.2에서 자세히 언급한다. 제안하는 방법에서 *t(a)*를 적용하는 이유는 직관적으로 높은 *Sim*을 갖는 CNVR과 유전자는 기능적으로 상관관계가 클 수 있기 때문이다. 지금까지 언급한 전체적인 수행과정은 (그림 3)과 같이 표현될 수 있다.

3. 실험 및 분석

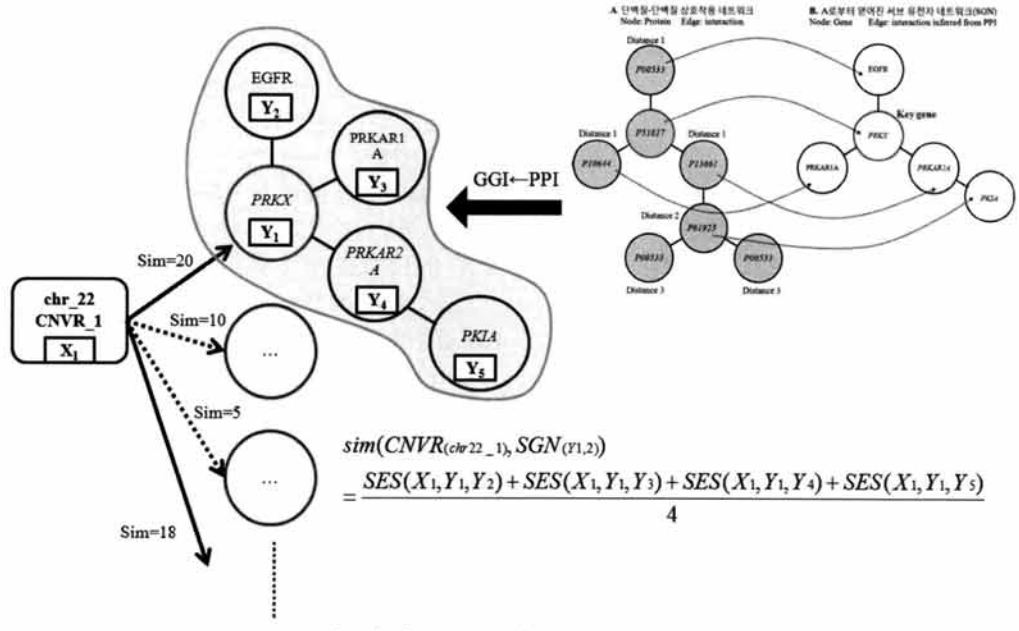
본 실험에서는 CNVR과 기능적으로 관련되어 있는 유의미한 유전자를 결정할 수 있도록 SES 결정을 위한 파라미터를 결정하는 방법을 제시하고, 그것을 만족하는 CNVR과 연관된 유전자들에 대해서 GO(Gene Ontology) 데이터베이스를 이용해서 기능적 연관성을 검증함으로써 결과를 검증한다.

3.1 실험 환경

제안하는 방법은 Standard Template Library를 이용한 C++로 구현되었으며, Windows 7 운영체제에서, 4GB의 메모리와 AMD Phenom™ II X2 545 Processor 3.0GHz 의 CPU를 가진 컴퓨터를 사용해 실험을 하였다.

3.2 파라미터 결정

본 장에서는 제안하는 방법에서 사용되는 파라미터인 *t(a)*을 결정하기 위한 방법을 기술한다. 본 논문에서 해결하고자 하는 문제에서는 가장 최적의 파라미터를 찾기 위해서 사용할 수 있는 정답집합이 존재하지 않기 때문에 *t(a)*를 결정하기 위해서는 무작위로 추출한 표본집단에서 유의수준 a에 해당하는 *Sim*값을 최적의 파라미터로 가정하였다. 이를 위해 전체 염색체에 대한 CNVR 데이터와 전체 SGN집합들에 대해서 10,000개의 무작위 표본 추출을 수행하였다. 즉

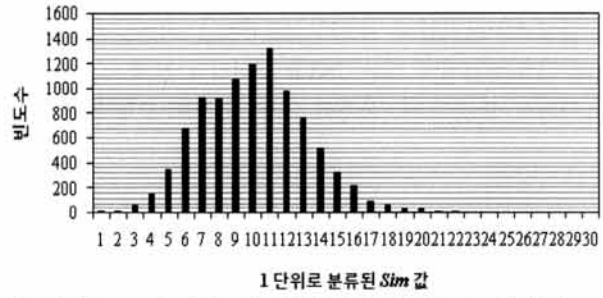


(그림 3) *Sim*을 구하는 전체 과정 도식도

무작위로 하나의 CNVR과 SGN를 추출하고 *Sim*를 구하는 과정을 총 10,000번 수행한다. 무작위 추출 후 얻은 *Sim*값의 분포는 (그림 4)와 같았다. 표본의 수가 30개 이상이기 때문에 정규분포를 따르는 것으로 가정할 수 있으며 α=0.05에 해당하는 *Sim*값을 파라미터로 결정하였다. α=0.05는 신뢰도 95%를 의미하며 표본 추출에 따른 $t(\alpha)$의 변화를 알아보기 위해서 이러한 표본추출 과정을 10번을 수행하였다. <표 2>는 무작위 표본 추출 실험에 대한 평균, 표준편차를 나타낸다. 또한 (그림 4)는 10번의 반복 실험 중 하나인 실험 2에 대해서 *Sim*값의 분포도를 나타내는데, 얻어진 분포도는 정규분포로 근사화될 수 있는 모습이었다. 10번의 반복 표본 추출 실험에서 얻은 *Sim*값의 평균과 표준편차는 거의 비슷하였고, 평균 t(0.05)인 16.41를 제안하는 방법의 *Sim* 파라미터로 사용하였다.

〈표 2〉 10번 반복한 무작위 표본 추출 실험에서 얻은 *Sim*값의 평균, 표준편차와 각 실험의 $t(0.05)$

No.	Mean	Standard Deviation	t(0.05)
1	9.67485	3.37187	16.4186
2	9.66194	3.35440	16.3707
3	9.60905	3.39683	16.4027
4	9.64330	3.36008	16.3635
5	9.63221	3.40178	16.4358
6	9.63027	3.40987	16.4500
7	9.64245	3.35972	16.3619
8	9.66086	3.40798	16.4768
9	9.71234	3.41482	16.5420
10	9.63093	3.34630	16.3235



(그림 4) 표 2의 실험 2의 무작위 추출에 대한 *Sim*값의 분포도

3.3 결정된 파라미터에 대한 검증

3.2에서 결정된 파라미터인 $t(0.05)$ = 16.41는 결과적으로 가능한 모든 유전자에 대해서 통계적으로 CNVR과 강하게 연관되어 있을 것이라고 판단된 유전자들만 선택할 수 있는 기준으로 쓰이게 된다. 본 논문에서 제안하고 있는 *Sim*은 하나의 CNVR에 대해서 기능적으로 연관성이 높은 유전자에 대해서 그렇지 않은 유전자보다 높은 값을 보인다. 이유는 기본적으로 CNVR과 유전자에 대해서 다수의 샘플에서 동일한 발현 패턴을 보인다면 기능적으로 높은 상관관계가 있다고 볼 수 있는데 본 논문에서 제안하고 있는 *Sim*은 $SGN_{(K,\ 2)}$에서 하위 유전자를 고려함으로써 PCC 같은 전통적인 상관관계 측정방법보다 노이즈를 고려하면서 높은 상관관계를 보이는 것들을 효과적으로 찾을 수 있기 때문이다. 따라서 하나의 CNVR에 대해서 $t(0.05)$ 이상을 갖는 유전자들은 기능적 관련성이 높다고 간주할 수 있다. 이것은 다시 말해서 해당 CNVR과 $t(0.05)$ 이상으로 선택된 유전자들은 $t(0.05)$를 고려하지 않고 임의로 뽑힌 유전자들 보다 CNVR과 기능적으로 강하게 연관되어 있을 것이고 동시에 유전자들끼리도 기능적 관련성의 정도가 클 수 있음을 유추할 수 있다. 예를 들어 하나의 CNVR이 A라는 유전자의 TF를 포함하고 있어서 A유전자의 발현에 영향을 미친다면 A유전자와 기능적으로 관련이 되어있는 다른 B유전자, C유전자 등에도 영향을 미칠 수 있게 된다. 따라서 그 CNVR과 A유전자와의 *Sim*과 B유전자와의 *Sim*, C유전자와의 *Sim* 모두 비슷하게 높은 값을 보일 가능성이 높다.

이런 가능성이 있을 수 있다는 것을 기본적으로 가정하면서 본 장에서는 실제로 하나의 CNVR과 기능적으로 연관되어 있다고 판단되는 유전자들의 집합, 즉 $t(0.05)$ 이상의 유전자들의 집합과, $t(0.05)$을 만족하지 않고 임의로 선택된 유전자들의 집합 사이에는 기능적인 차이가 존재함을 실험을 통하여 밝히고자 했다. 유전자의 기능적 관련성은, 유전자의 기능을 생물학적으로 밝혀서 데이터베이스화 한 GO (Gene Ontology) 데이터베이스를 이용해서 측정할 수 있으며, 여기서는 유전자 집합의 기능을 GO 데이터베이스를 이용해서 실시하는 검증을 GO 검증이라고 한다. 즉, $t(0.05)$를 적용하지 않고 임의로 선택된 유전자들에 대한 GO검증 결과와 $t(0.05)$를 적용했을 경우의 GO검증 결과의 차이를 통하여 우리가 결정한 파라미터의 유의성을 증명할 수 있었다. GO 검증은 FuncAssociate2.0 [18] 의 클라이언트 모델을 이용하여 수행되었다.

우리는 각 염색체의 각 CNVR에 대해서 $t(0.05)$을 적용했을 경우 얻을 수 있는 키유전자들과 함께 비교대상으로 동일한 개수의 키유전자들을 임의로 선택하였을 경우의 GO 검증에서 얻은 p-value의 차이를 알아보았다. <표 3>은 22번 염색체에 대한 실험 결과이다. 염색체 22번의 경우 총 2개의 CNVR 가운데 $t(0.05)$을 적용한 경우 11개의 CNVR에서 GO검증이 완료되었다. 하지만 임의로 동일 개수만큼 유전자를 선택한 경우는 총 2개의 CNVR에 대해서만 GO검증이 수행되었다. 전체적으로 보았을 경우 $t(0.05)$를 적용했을 경우 GO검증이 더 잘 이루어졌다고 볼 수 있다. 또한 얻어진 p-value 또한 대체로 낮은 값을 가지고 있기 때문에 GO 검증이 이루어진 CNVR과 연결된 유전자들은 기능적으로 높은 상관관계를 갖고 있다고 말할 수 있다.

<표 3과 4>에서 보는 바와 같이 모든 CNVR에 대해서 GO 검증이 이루어지지는 않았으며, 검증된 CNVR에 대해서도 모두 $t(0.05)$을 적용한 편이 더 낮은 p-value를 보이지는 않았다. 하지만 대략적인 추세는 확실히 $t(0.05)$를 적용한 것과 그렇지 않은 것은 p-value의 차이를 보였으며 p-value의 차이가 큰 CNVR에 대해서는 해당 유전자들과 기능적으로 강한 연관성이 있다는 것을 알 수 있었다. 모든 염색체에 대해서 실험결과를 정리하기에는 차이가 있었기에 대표적으로 22번 염색체와 가장 크기가 큰 1번 염색체에 대해서 결과를 <표 3, 4>에 보였다.

〈표 3〉 22번 염색체에 대하여 모든 CNVR에 대한 GO 검증 결과

No. Chromosome 22	P-value^T	P-value^R	No. of found genes with $t(0.05)$	GO Term
CNVR_1	NA	NA	16	-
CNVR_2	NA	NA	17	-
CNVR_3	NA	NA	3	-
CNVR_4	NA	0.02	593	GO:0048523
CNVR_5	0.014	NA	84	GO:0070161
CNVR_6	0.012	0.003	9546	GO:0043681/GO:0005829
CNVR_7	NA	NA	12	-
CNVR_8	0.002	NA	90	GO:0042611
CNVR_9	0.005	NA	124	GO:0005576
CNVR_10	0.001	NA	49	GO:0005576
CNVR_11	NA	NA	450	-
CNVR_12	NA	NA	1	-
CNVR_13	NA	NA	3	-
CNVR_14	NA	NA	84	-
CNVR_15	NA	NA	63	-
CNVR_16	0.006	NA	97	GO:0005587
CNVR_17	NA	NA	196	-
CNVR_18	< 0.001	NA	133	GO:0070161
CNVR_19	0.001	NA	33	GO:0042613
CNVR_20	0.018	NA	225	GO:0030057
CNVR_21	< 0.001	NA	34	GO:0005576
CNVR_22	< 0.001	NA	43	GO:0005576
CNVR_23	NA	NA	3	-
CNVR_24	NA	NA	2	-
CNVR_25	NA	NA	12	-
CNVR_26	NA	NA	0	-

* N/A는 SGN에 대해서 GO 테스트를 수행했지만 결과를 얻지 못한 경우를 나타냄
* P-value^T 는 t(0.05)를 적용한 경우 유전자들에 대한 GO검증에서 얻어진 adjusted P-value 결과를 나타냄
* P-value^R 은 동일 개수만큼 임의로 선택된 유전자들에 대한 GO검증에서 얻은 adjusted P-value를 나타냄

〈표 4〉 1번 염색체에 대하여 모든 CNVR에 대한 GO 검증 결과

No. Chromosome 1	P-value^T	P-value^R	No. of found genes with $t(0.05)$	GO Term
CNVR_1	N/A	N/A	33	-
CNVR_2	N/A	N/A	83	-
CNVR_3	< 0.001	N/A	165	GO:0005615
CNVR_4	0.003	N/A	168	GO:0005615
CNVR_5	N/A	N/A	15	-
CNVR_6	0.016	N/A	134	GO:0016337
CNVR_7	N/A	N/A	3	-
CNVR_8	0.009	N/A	342	GO:0016525
CNVR_9	N/A	N/A	0	-
CNVR_10	N/A	N/A	0	-
CNVR_11	< 0.001	N/A	8	GO:0042611
CNVR_12	N/A	N/A	1	-
CNVR_13	< 0.001	N/A	274	GO:0016337
CNVR_14	N/A	N/A	1	-
CNVR_15	N/A	0.005	56	GO:0032991
CNVR_16	N/A	N/A	5	-
CNVR_17	N/A	N/A	58	-
CNVR_18	N/A	N/A	1	-
CNVR_19	0.005	0.016	169	GO:0005911/GO:0051014
CNVR_20	N/A	N/A	42	-
CNVR_21	0.002	N/A	79	GO:0016337
CNVR_22	0.007	N/A	235	GO:0005911
CNVR_23	N/A	N/A	6	-
CNVR_24	0.006	0.004	2147	GO:0006626/GO:0005829
CNVR_25	N/A	N/A	85	-
CNVR_26	N/A	N/A	0	-
CNVR_27	0.002	N/A	125	GO:0005576
CNVR_28	0.001	N/A	239	GO:0005576
CNVR_29	0.002	N/A	29	GO:0005576
CNVR_30	N/A	N/A	0	-

CNVR_31	N/A	N/A	431	-
CNVR_32	0.002	N/A	125	GO:0042613
CNVR_33	0.012	N/A	9	GO:0008329
CNVR_34	0.012	N/A	29	GO:0030057
CNVR_35	0.001	N/A	389	GO:0005201
CNVR_36	N/A	N/A	2	-
CNVR_37	N/A	N/A	32	-
CNVR_38	N/A	N/A	491	-
CNVR_39	N/A	N/A	0	-
CNVR_40	0.001	N/A	43	GO:0005576
CNVR_41	N/A	N/A	23	-
CNVR_42	< 0.001	0.008	999	GO:0006626/GO:0005515
CNVR_43	N/A	N/A	0	-
CNVR_44	N/A	N/A	4	-
CNVR_45	N/A	N/A	3	-
CNVR_46	N/A	N/A	9	-
CNVR_47	N/A	N/A	213	-
CNVR_48	0.007	0.002	726	GO:0019843/GO:0005515
CNVR_49	< 0.001	N/A	49	GO:0005576
CNVR_50	N/A	N/A	0	-
CNVR_51	N/A	N/A	0	-
CNVR_52	0.007	N/A	184	GO:0004867
CNVR_53	N/A	N/A	1	-
CNVR_54	0.009	N/A	160	GO:0005911
CNVR_55	N/A	N/A	17	-
CNVR_56	N/A	N/A	14	-
CNVR_57	N/A	N/A	5	-
CNVR_58	N/A	N/A	66	-
CNVR_59	0.011	N/A	45	GO:0016337
CNVR_60	0.012	N/A	121	GO:0005587
CNVR_61	< 0.001	N/A	45	GO:0005576
CNVR_62	N/A	N/A	12	-
CNVR_63	0.003	N/A	204	GO:0016337
CNVR_64	N/A	N/A	255	-
CNVR_65	N/A	N/A	175	-
CNVR_66	N/A	N/A	1	-
CNVR_67	N/A	N/A	60	-

3.4 하나의 CNVR과 각 키유전자 사이의 연관성 분석 결과

3.3에서는 하나의 CNVR에 대해서 $t(0.05)$을 만족하는 모든 키유전자들을 대상으로 GO검증을 수행하였다. 마지막으로 본 실험에서는 낮은 P-value를 갖는 하나의 CNVR에 대해서 선택된 각각의 키유전자 K에 해당하는 $SGN_{(K,\ 2)}$에 대해서 GO 검증을 수행하였다. 이 실험을 통해 우리는 하나의 CNVR과 연관되어 있을 가능성이 큰 유전자기능들의 후보를 더 많이 얻을 수 있었다. <표 5>는 대표적으로 1번 염색체의 11번째 CNVR에 대하여 분석을 수행한 결과를 나타낸다. 총 10638개의 가능한 SGN들에서 1번 염색체의 CNVR[16509639-17229287]에 대해서 $t(0.05)$이상의 SES값을 갖는 SGN는 8개였다. 이중 GO 검증 결과 P-value를 얻은 것은 5개였다. 이 중에서 특히 가장 낮은 P-value를 갖는 1번 SGN는 CNVR[16509639-17229287]과 강하게 관련이 되어있을 가능성이 높다고 할 수 있다. 전반적으로 이 실험을 통하여 해당 CNVR은 GO:0030101(natural killer cell activation, 세포 자살), GO:0005923(tight junction, 세포간 결합), GO:0042613(MHC class II protein complex, 항원 복합체), GO:0006874(cellular calcium ion homeostasis, 칼슘 이온 항상성), GO:0031093(platelet alpha granule lumen, 혈소판 성장 요소 소기관) 과 같은 기능의 발현에 전반적으로 연관성이 있으며, 이를 통해서 1번 염색체의 11번째 CNVR은 전반적으로 세포의 유지에 관여하고 있다는 것을 추론해 볼 수 있다. 본 논문에서는 CNVR과 연관된 유전자 후보군을 시스템적으로 제시하고, 그와 함께 얻어진 후보 유전자집단에 대한 GO검증을 통해서 CNVR의 기능을 추론하고 있다.

이와 같은 연관성 분석 결과는 기존에는 이루어지지 않았던 새로운 형태의 실험 결과이다. 본 논문에서 사용한 데이터와 동일한 데이터를 사용한 [14]에서는 하나의 CNVR에 대해서 시작과 끝 위치의 2Mb이내에 존재하는 유전자들에 대해서만 연관성 분석을 수행하였다. <표 6>은 [14]의 최종 결과의 일부분을 나타낸다. 예를 들어 첫 번째 행인 Chr19_9는 해당 CNVR의 시작점으로부터 514,966위치와 끝 점으로부터 682,687이내에 유전자 "GI_11038644-I"가 존재하며 연관성을 정도를 나타내는 p-value는 -log 스케일 후 3.5605이다. <표 6>의 모든 행은 0.001 이상의 p-value를 나타내는 결과만 모아놓은 것이다. [14]의 결과는 하나의 CNVR에 대해서 특정 영역 내에 존재하는 유전자만을 대상으로 연관성을 검사하였기 때문에, <표 5>와 같이 하나의 CNVR에 대해서 GO검증을 하지 못할 뿐만 아니라 특정 기능에 대한 유추도 사실상 어려웠다. [14] 이후 사실상 정상인 유전체 집단에 대해서는 연관성 연구가 이루어지지 않았

〈표 5〉 FuncAssociate 2.0을 이용한 1번 염색체의 11 번째 CNVR[16509639-17229287] 에 대한 GO 검증 결과

SGN id	SES	N	X	Adjusted P-value	GO Term	SGN
1	16.6667	2	24	< 0.001	GO:0030101	GI_13375655-S GI_6679051-S
2	20.7143	N/A	N/A	N/A	N/A	GI_16418366-S GI_20149581-S GI_4506922-S
3	17.1429	N/A	N/A	N/A	N/A	GI_21040248-S GI_6679051-S
4	22.381	2	71	0.004	GO:0005923	GI_21536296-S GI_28416401-A
5	16.6667	2	15	0.002	GO:0042613	GI_24797075-S GI_24797073-S
6	17.1429	N/A	N/A	N/A	N/A	GI_31341185-S GI_6226959-S
7	20	2	146	0.007	GO:0006874	GI_4506832-S GI_13929430-S
8	17.619	2	35	0.002	GO:0031093	GI_45269140-S GI_10518500-S

* N/A는 SGN에 대해서 GO 테스트를 수행했지만 결과를 얻지 못한 경우를 나타냄
* 'SGN(Gene Set)' 열에서 키유전자는 굵게 표시
* N은 유전자 집합 중에서 GO에 속하는 유전자의 개수를 나타냄
* X는 GO를 이루는 유전자의 개수를 나타냄

〈표 6〉 [14]에서 구한 CNVR과 gene expression 사이의 연관성검사 결과의 일부분

CNV_ID	Illumina Gene_ID	-log10 (P-value)	Distance probe to CNV_start	Distance probe to CNV_end	CNV_class	CNV Frequency in 270 HapMap individuals
Chr19_9	GI_11038644-I	3.5605	514,966	682,687	del	3
Chr10_6	GI_31341468-S	3.2798	1,775,769	1,931,816	del	54
Chr22_26	GI_25092724-S	4.2056	-7,936	93,355	dup	25
Chr3_25	GI_27894375-S	6.1255	-90,713	93,983	complex	3
Chr2_10	GI_7656998-S	7.1119	-212,771	-33,218	del	2
Chr22_24	GI_32698823-S	6.9669	-95,275	73,303	del/dup	27
Chr14_6	GI_31742485-S	3.585	-1,535,518	-1,377,556	dup	4

고, [14]에서 수행한 방법과 같이 단순히 두 데이터 사이의 선형적인 상관계수를 측정하는 방법이 대부분이었기 때문에 제안하는 방법의 우수성을 입증하기 다른 알고리즘들과 비교 실험을 수행하는 것은 큰 의미는 없었다. 본 논문의 목표가 CNVR의 기능을 효과적으로 분석할 수 있도록 이질 생물 데이터를 통합하고 새로운 연관성 측정방법을 사용하는 분석 모델을 제안하는 것이었기 때문에 <표 5>와 같은 새로운 분석 결과를 도출해 낸 것이 관련 분야에 대한 본 논문의 가장 큰 공헌이라고 할 수 있다.

4. 결 론

본 논문에서는 CNVR의 기능적 특성을 cis-유전자들 뿐만 아니라 trans-유전자 수준에서도 밝힐 수 있는 새로운 연관성 분석 방법론을 제안하였다. 본 연구의 공헌은 다음과 같다. (1) 정상인 유전체에 대해서 trans-유전자와의 기능적 상관도를 유추할 수 있도록 다양한 이질 생물학 데이터에 대한 통합 방법론을 제시하였다. (2) 제안하는 방법에 효과적인 새로운 연관성 측정 방식인 $SES(X_1, Y_1, Y_2)$를 제시하였다. (3) GO검증을 통하여 $SES(X_1, Y_1, Y_2)$를 통해서 얻은 Sim의 유의성을 보였으며, CNVR과 관계된 유전적 기능의 후보들을 자동적으로 제시할 수 있는 시스템을 제안하였다.

현재까지 수행된 정상인 기반의 CNVR과 유전자 발현과의 연관성 연구는 CNVR에 대해서 근거리의 유전자들만 대상으로 수행하였기 때문에 찾아낸 결과는 한계가 있었다. 하지만 제안된 방법을 이용하면 CNVR과 연관된 유전적 기능 후보들을 높은 유의성과 함께 제시할 수 있으며, 시스템적으로 모든 분석 방법이 구현되어있기 때문에 후보 결과들을 쉽게 얻을 수 있다. 향후 본 논문에서는 새롭게 출시되고 있는 고해상도 aCGH 데이터와 최근 이슈가 되고 있는 유전자-유전자 상호작용(GGI) 데이터[19]를 이용하여 CNVR과 연관성을 분석할 수 있는 방법론에 대한 연구를 수행할 예정이다.

참 고 문 헌

[1] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," Nature, Vol.409, No.15, pp.860-921, 2001.

[2] J. Hampe, A. Franke, et al., "A genomewide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1," Nat. Genet., Vol.39, No.2, pp.207-211, 2007.

[3] CI. Amos, et al., "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.," Nat. Genet. Vol.40, pp.616-622, 2008.

[4] G.M. Cooper, et al., "Systematic assessment of CNV detection via genome-wide SNP genotyping," Nat. Genet., Vol.50, pp.1199-1203, 2008.

[5] J. Beckmann et al., "Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic

variability," Nat. Genet., Vol.8, No.8, pp.639-646, 2007.
[6] S.A. McCarroll, and D.M. Altshuler, "Copy-number variation and association studies of human disease," Nat. Genet., Vol.39, pp.S37 - S42, 2007.
[7] R. Redon, et al., "Global variation in copy number in the human genome," Nature, Vol.444, pp.444 - 454, 2006.
[8] E. Tuzun, et al., "Fine-scale structural variation of the human genome," Nat. Genet., Vol.37, No.7, pp.727-732, 2005.
[9] 김태민, "Copy number variation (CNV)의 인간유전체 내 기능 및 진화경로에 관한 연구," 생화학분자생물학소식, Vol.15, No.3, pp.40-51, 2008.
[10] H. Lee, et al., "Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes," Bioinformatics, Vol.24, No.7, pp.889-896, 2008.
[11] S. Junnila, et al., "Genome-wide gene copy number and expression analysis of primary gastric tumors and gastric cancer cell lines," BMC Cancer, Vol.10, No.73, 2010.
[12] The Wellcome Trust Case Control Consortium, "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls," Nature, Vol.464, No.7289, pp.713-720, 2010.
[13] S.A. McCarroll, "Copy-number analysis goes more than skin deep," Nat. Genet., Vol.40, pp.5 - 6, 2008.
[14] B.E. Stranger, et al., "Relative impact of nucleotide and copy number variation on gene expression phenotypes," Science, Vol.315, pp.848 - 853, 2007.
[15] D.F. Conard, et al., "Origins and functional impact of copy number variation in the human genome", Nature, Vol.464, No.7289, pp.704-712, 2010.
[16] The International HapMap Consortium, "A haplotype map of the human genome," Nature, Vol.437, No.7063, pp.1299-1320, 2005.
[17] K.R. Brown, I., Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks," Genome Biology, Vol.8, No.5, pp.R95, 2007.
[18] G.F. Berriz, et al., "Next-generation software for functional trend analysis," Bioinformatics, Vol.25, No.22, pp.3043-3044, 2009.
[19] A. Battle, M. C Jonikas, et al., "Automated identification of pathways from quantitative genetic interaction data," Molecular Systems biology, Vol.6. No.379, 2010.
[20] 황기태, 정중기, 외 8명, "Array CGH를 이용한 DNA 복제 수 변화에 대한 분석에 따른 유방암의 예후 인자에 대한 후보 유전자로서의 COL18A1," Journal of breast cancer, Vol.13, No.1, pp.37-45, 2010.
[21] C.N. Henrichsen, et al., "Copy number variants, diseases and gene expression," Human Molecular Genetics, Vol.18, Review Issue.1, pp.R1-R8, 2009.
[22] J.O. Korbel, et al., "The current excitement about copy-number variation: how it relates to gene duplications and protein families," Current Opinion in Structural Biology, Vol.18, pp.366-374, 2008.
[23] Wellcome Trust Sanger Institute: http://www.sanger.ac.uk/humgen/cnv/redon2006/
[24] Interologous Interaction Database: http://ophid.utoronto.ca/phidv2.201/

박 치 현

e-mail : tianell@cs.yonsei.ac.kr
2007년 홍익대학교 컴퓨터공학과(학사)
2009년 연세대학교 컴퓨터과학과 (공학석사)
2009년~현 재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 바이오인포매틱스, 데이터마이닝, 데이터베이스시스템

안 재 균

e-mail : ajk@cs.yonsei.ac.kr
2006년 연세대학교 컴퓨터과학과(학사)
2009년 연세대학교 컴퓨터과학과 (공학석사)
2009년~현 재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스

윤 영 미

e-mail : ymyoon@gachon.ac.kr
1981년 서울대학교 자연과학대학(학사)
1983년 오하이오주립대학교 수학과 (학사수료)
1987년 스탠포드대학교 컴퓨터과학과 (이학석사)
2008년 연세대학교 컴퓨터과학과(공학박사)
1987년~1993년 IntelliGenetics Inc., California, USA, Software Engineer
1995년~현 재 가천의과학대학교 정보공학부 교수
관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스

박 상 현

e-mail : sanghyun@cs.yonsei.ac.kr
1989년 서울대학교 컴퓨터공학과(학사)
1991년 서울대학교 컴퓨터공학과 (공학석사)
2001년 UCLA 대학원 컴퓨터과학과 (공학박사)
1991년~1996년 대우통신 연구원
2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow
2002년~2003년 포항공과대학교 컴퓨터공학과 조교수
2003년~2006년 연세대학교 컴퓨터과학과 조교수
2006년~현 재 연세대학교 컴퓨터과학과 부교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오인포매틱스, 적응적 저장장치 시스템, 플래쉬메모리 인덱스, SSD