암 연구를 위한 데이터 및 텍스트 마이닝

Data and Text Mining for Cancer Research

저자
(Authors)

장호, 김정균, 이현주
Ho Jang, Jeongkyun Kim, Hyunju Lee

출처
(Source)

정보과학회지 32(3), 2014.3, 61-70 (10 pages)
Communications of the Korean Institute of Information Scientists and Engineers 32(3), 2014.3, 61-70 (10 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE02373853

APA Style

장호, 김정균, 이현주 (2014). 암 연구를 위한 데이터 및 텍스트 마이닝. 정보과학회지, 32(3), 61-70.

암 연구를 위한 데이터 및 텍스트 마이닝

광주과학기술원 ▌장 호 · 김정균 · 이현주

1. 서 론

인간 세포내의 DNA, RNA, 단백질 등의 분자는 암의 발병 및 진행과 함께 변화를 나타낸다. DNA의 염기 서열(Nucleotides Sequence)에서는 점 변이(Point Mutation), 구조적 변이(Structural Variation), DNA 유전자 복제수 변이(DNA Copy Number Aberrations, CNAs) 등의 변화가 일어난다. 이에 따라서 DNA로부터 전사되는 mRNA 및 microRNA 등의 발현량(Expression Level) 및 단백질의 구조 등이 변화하게 된다(그림 1)[1]. 이러한 DNA, RNA, 단백질 등의 서로 다른 분자 형태에서 다양하게 일어나는 암 세포의 변화는 분자들이 상호 연관성을 가지고 서로 영향을 미치기 때문에 더욱 복잡하다. 유사한 기능을 가진 다수의 유전자들은 함께 생물학적 경로(Biological Pathway)에서 활동하는데, 암세포에서는 유전자들의 변화에 의해서 세포 주기(Cell Cycle), 세포 증식(Cell Proliferation), 세포 괴사(Apoptosis) 등에 관련된 생물학적 경로가 변화하게 된다(그림 2)[2]. 따라서, 암과 관련된 유전자를 발굴하는 연구와 함께, 암의 분자 기전에 대한 연구 또한 중요한 연구 주제이다. 유전자 및 생물학적 경로의 변화는 수백 개의 암의 종류(간암, 대장암, 유방암, 난소암, 자궁암 등) 및 그 세부 암(Cancer Subtype) 종류에 따라서 특이적인 변화를 가지기 때문에, 이 특이적인 변화들에 대한 연구가 활발히 진행되고 있다[3]. 암의 분자 기전에 대한 연구는 최근 The Cancer Genome Atlas[3] 및 Cancer Cell Line Encyclopedia[4] 등에서 발표한 대용량 암 게놈 관련 데이터에 의해서 가속화되고 있다. 2014년 2월 현재 cBioPortal[5] 웹 서비스에서는 56개의 암 게놈 연구에서 발표한 15,506개의 암환자 샘플에 대한 데이터를 제공하고 있다.

† 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013058053).

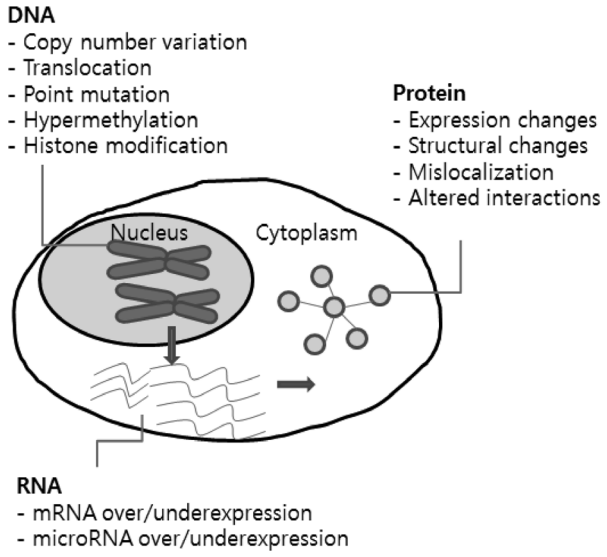

그림 1 암 세포에서의 분자들의 변화

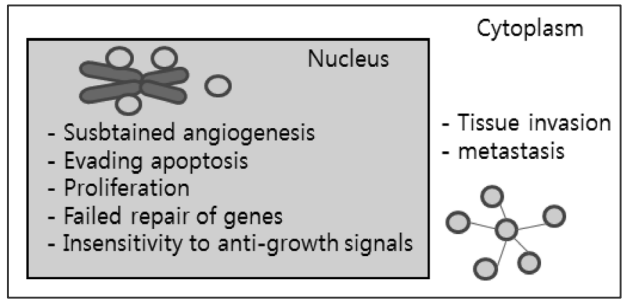

그림 2 암 세포에서의 생물학적 경로의 변화

생물정보학에서는 암의 발생 및 진행과 관련된 유전자 및 생물학적 경로를 계산학적으로 예측하는 연구를 활발히 해 오고 있다. 연구는 크게 분자 레벨의 데이터를 분석하는 데이터 마이닝 연구와 생의학 논문에 보고된 기존 연구들을 분석하는 텍스트 마이닝 연구로 나눌 수 있다. 본 논문의 2장에서는 암과 관련된 데이터 마이닝 연구의 한 분야로써, DNA 유전자 개수 이상의 변화 데이터를 이용하여, 암과 관련된 유전자를 발굴하는 다양한 방법론들에 대해서 소개한다.

그 후 CNAs, 유전자 발현량(Gene Expression, GE) 및 단백질 상호 작용(Protein-Protein Interactions, PPI)을 통합하는 알고리즘을 설명하고, 이를 대용량 암환자 데이터에 적용한 연구들을 소개한다. 3장에서는 생의학 문헌에 발표된 암과 관련된 유전자를 체계적으로 분석하여, 각종 암의 타입에 따라 중요한 유전자들을 우선순위별로 정렬해 문헌 정보와 함께 보고하는 텍스트 마이닝 기법에 대해서 설명한다.

2. 암 연구를 위한 데이터 마이닝

2.1 유전변이의 개념과 종류

바이오테크놀로지의 발전으로 대량의 인간 유전체 데이터가 축적되어 질병과 유전변이(Genetic Variation) 사이의 연관성에 대한 연구 결과들이 보고되고 있다. 일례로 염색체 19번의 유전자 ApoE의 일부 염기만 바뀌어도 알츠하이머병에 걸릴 위험이 몇 배나 높아진다고 알려져 있다[6]. 유전변이란 집단내의 또는 집단 간의 대립 유전자(Allele)의 변이를 의미한다. 대립유전자란 쌍으로 이루어진 대립형질의 유전자를 뜻한다. 유전변이는 종류에 따라 그림 3과 같이 점 돌연변이와 구조적 변이로 나눌 수 있다. 점 돌연변이는 염색체 상의 하나의 염기가 다른 값으로 변환되거나 추가 또는 제거되는 형태의 돌변변이이다. CNAs는 염색체의 한 단편이 중복(Gain)되거나 결실(Deletion) 되는 형태로 나타난다. 구조적 변이는 CNAs를 포함하는 개념으로 염색체의 특정 단편이 뒤집어진 역위(Inversion)와 염색체의 단편이 같은 염색체의 다른 위치로 또는 다른 염색체로 이동한 전좌(Translocation)가 있다.

2.2 유전자 복제수 변이 검출의 중요성

CNAs는 유전 변이의 주요한 한 종류로 표 1과 같은 다양한 질병과의 연관성이 보고되었다. 특히 CNAs는

표 1 CNAs와 연관성이 보고된 질병들

질병	연도	저널
알츠하이머병(Early-onset Alzheimer's disease)[8]	2006	Nat Genet
자폐증(Autism)[9]	2007	Science
자가면역질환(Autoimmune Disease)[10]	2007	Nat Genet
암(Cancer)[11]	2008	Nat Genet
정신분열증(Schizophrenia)[12]	2008	Nature
비만(Obesity)[13]	2010	Nature

암 세포와 큰 연관성이 있다[7]. 발암(Carcinogenesis)이 진행 중인 염색체는 구조적인 변화를 겪고 이는 암 유발 유전자(Oncogene) 또는 암 억제 유전자(Tumor Suppressor Gene)의 유전자 복제수에 변화를 불러올 수 있다. 암 유발 유전자의 유전자 복제수 증가나 암 억제 유전자의 유전자 복제수 감소는 암의 발생과 진행에 큰 영향을 미칠 수 있기 때문에 암과 CNAs 사이의 관계를 올바르게 파악하는 것은 암의 연구에 있어 핵심적이다. 이를 위해 염색체 상에 CNAs가 일어난 위치를 밝히고 그 영역이 중복(Copy Number Gain) 되었는지 또는 제거(Copy Number Deletion) 되었는지 추정하는 것이 매우 중요하다.

2.3 CNAs 검출을 위한 바이오테크놀로지

CNAs를 검출하기 위해 사용되는 일반적인 기술은 비교유전체보합법(Comparative Genomic Hybridization, CGH)이나 SNP 마이크로어레이이다. 이 기술은 DNA가 상보적인 가닥에 붙는 원리를 이용한 것으로 널리 사용되고 있다. 하지만, 이러한 플랫폼의 경우 최대 수백만개 정도의 프로브(Probe)를 가지므로 30억 개의 염기쌍을 가지는 인간 유전체를 정확하게 측정하기에는 한계가 있다. CNAs를 찾는 또 다른 방법으로는 DNA 시퀀싱 데이터를 활용하는 방법이 있다. 차세대 염기서열 분석(Next Generation Sequencing, NGS) 기술의 등장

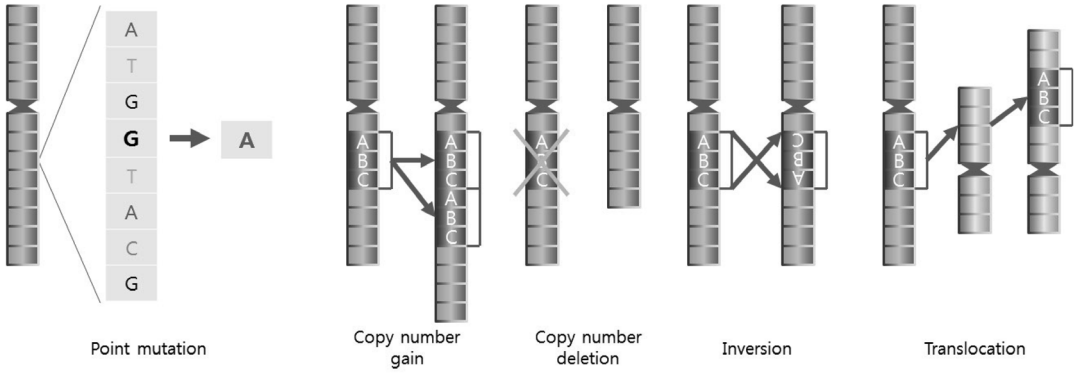

그림 3 유전변이의 종류

이후 시퀀싱 비용이 큰 폭으로 감소했기 때문에 최근에 시퀀싱 데이터를 이용한 CNAs 영역 검출 알고리즘들이 꾸준히 발표되고 있다. 시퀀싱 데이터를 이용한 CNAs 검출 알고리즘은 마이크로어레이 데이터를 이용한 알고리즘 보다 더욱 정확한 영역 측정과 유전자 복제수의 추정이 가능하다.

2.4 개별 샘플의 CNAs 검출 알고리즘

개별 샘플 CNAs를 검출하기 위해서는 먼저 조직에서 세포를 추출하고 마이크로어레이나 NGS를 이용하여 분석 가능한 데이터를 생성한다. 그러나 이러한 데이터에는 노이즈가 포함되어있어 육안으로 CNAs 영역을 정확하게 알아내기는 쉽지 않다. 예를 들어 암세포의 CNAs를 연구하기 위해서는 암세포로부터 DNA를 추출해야 하는데 이 과정에서 정상세포에 포함된 DNA가 일정 비율 혼합될 수 있다. 그 밖의 다른 여러 노이즈로 인해 실험에서 얻어진 염색체의 복제수는 그림 4의 점과 같이 노이즈가 포함된 값으로 표현된다. 따라서, 이러한 노이즈를 적절히 처리하여 CNAs 영역을 검출하기위한 세그멘테이션 알고리즘들이 개발되었다. 그림 4의 직선은 알고리즘을 적용한 후의 복제수를 나타낸다. 세그먼트에서 유전자 복제수 2인 경우는 변화가 없는 영역, 3인 경우는 유전자가 한 번 더 복제된 영역을 뜻하고, 1인 경우는 두 상동 염색체 중 한 염색체에 결손이 일어난 것을 의미한다.

마이크로어레이 데이터를 사용하여 CNAs를 검출하기 위한 알고리즘으로 Circular Binary Segmentation (CBS)과 Gaussian-based-likelihood Approach(GLAD)가 널리 사용된다[14,15]. CBS 알고리즘은 체인지 포인트(Change Point)라는 개념을 사용하는데 이는 염색체 상의 유전자 복제수가 변하는 지점을 뜻한다. 염색체 상의 특정 영역에 체인지 포인트가 없다는 귀무가설과 그렇지 않다는 대립가설 사이의 통계량을 검정하여 통계적으로 유의미할 경우 해당영역에 CNAs가 있다고 판단한다. 예를 들어 그림 5에서 CBS는 염색체상의 60에서 160 사이의 영역과 그 외의 다른 영역사이에 체인지 포인트의 유무에 대한 통계검정을 수행한다. 그림 5는 60에서 160 사이의 영역에 CNAs가 있음을 나타낸다.

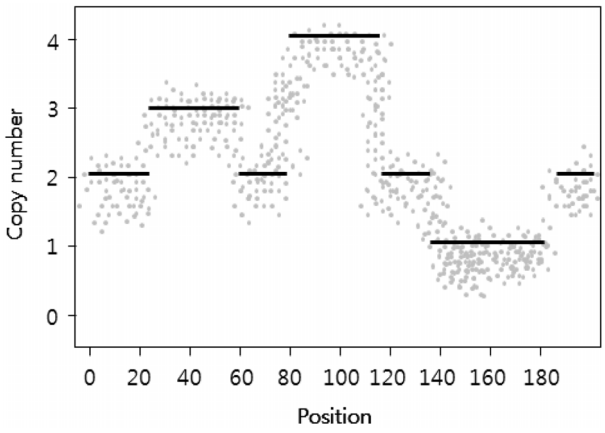

그림 4 개별 샘플의 유전자 복제수 변이 검출

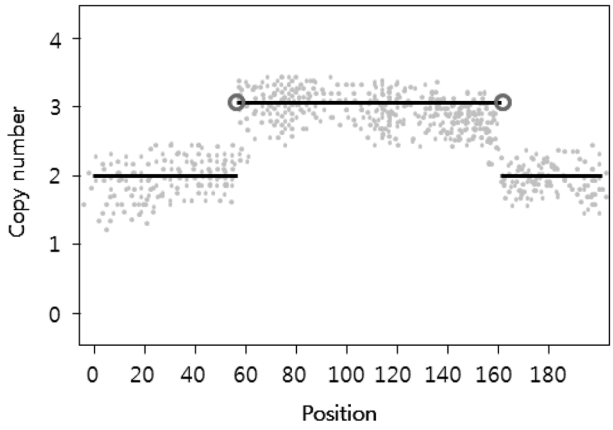

그림 5 Circular Binary Segmentation 알고리즘

2.5 주도적 변이 검출을 위한 알고리즘

인간의 유전체에서 CNAs가 차지하는 비율이 약 12%라고 한다[16]. 그러나, 모든 유전변이가 질병과 관련이 있는 것은 아니다. 마찬가지로 우리가 CNAs 검출 알고리즘을 이용하여 개별 샘플에서 일어나는 다수의 CNAs 영역을 발견했을 때 발견된 모든 유전변이가 암과 관련 있는 것은 아니다. 암의 발생 및 진행과 관련이 깊은 유전변이를 주도적 변이(Driver Mutation)이라고 부르고 암의 발생과는 무관한 유전변이를 승객변이(Passenger Mutation)이라고 부른다. 다수의 샘플에서 공통으로 CNAs가 일어나는 염색체 상의 영역을 찾음으로써 암을 유발할 가능성이 있는 주도적 변이를 발견할 수 있다.

GISTIC(Genomic Identification of Significant Targets in Cancer) 알고리즘은 다수의 샘플에 대한 CNAs를 조사함으로써 주도적 변이를 찾기 위한 알고리즘으로 널리 사용된다[17]. GISTIC 알고리즘의 대략적인 절차는 다음과 같다. 1단계에서 개별 샘플에 대해 유전 변이를 검출한 후 2단계에서 모든 영역의 유전자 복제수를 합한다. 이 값을 G-score라고 한다. 한편으로 퍼뮤테이션 테스트를 통해 임의로 일어나지 않을 G-score 임계치를 계산한다. 3단계에서는 G-score 임계치보다 적은 G-score 값을 가지는 영역을 제거한다. 오직 G-score 임계치 보다 큰 영역만 주도적 변이 영역으로 제안된다.

주도적 변이를 찾기 위해 웨이블릿을 이용한 연구

로는 WIFA(Wavelet-based Identification of Focal Genomic Aberrations)[18]가 있다. 웨이블릿은 신호의 노이즈를 제거하기 위한 강력한 도구로 널리 쓰이고 있다. 일반적으로 신호의 노이즈를 제거하려면 웨이블릿 변환(Wavelet Transform)을 통해 웨이블릿 계수와 스케일링(Scaling) 계수로 신호를 변환하고 웨이블릿 계수에 잡음을 제거한 후 다시 역변환(Inverse Transform)하여 노이즈가 제거된 신호를 복원하는 방식으로 사용된다. WIFA알고리즘은 다수의 샘플로부터 주도적 변이를 추정하기 위해 그림 6과 같은 단계로 진행된다. 1단계에서 개별 샘플의 유전변이가 입력되면 웨이블릿 변환을 통해 웨이블릿 계수와 스케일링 계수로 분해한다. 웨이블릿 계수만을 이용해 역변환을 하면 국소적인 변화(focal aberrations)를 나타내는 2단계와 같은 데이터가 생성된다. 이 신호는 유전자 복제수의 영역이 변화한 지점의 경계 값을 나타낸다. 이러한 변환을 모든 샘플에 대해 수행한 후 매트릭스로 표한 것이 3단계의 그림이다. 4단계에서 각 영역의 값들을 모두 더한 후 5단계에서 동일한 복제수 변이 패턴을 보이는 영역을 군집화 하는데, 이 군집들이 주도적 변이 영역을 나타낸다. WIFA 알고리즘을 154개의 교모세포종 뇌암(Glioblastoma multiforme, GBM) 샘플에 적용했을 때, 알려진 대부분의 GBM 유전자를 발견할 수 있었다(6단계). CDKN2A, PTEN, RB1의 경우 유전자 복제수가 감소하였고 나머지 유전자의 경우 유전자 복제수가 증가하였다.

2.6 암 연구를 위한 CNAs와 GE의 통합 방법론

암세포 속의 DNA의 특정 영역의 증가 및 감소에 따라서 그 영역 안에 포함된 유전자의 발현량이 어떻게 변화하는가에 관한 다양한 연구들이 진행되었다[19]. 최근 연구에서는 CNAs에 의한 GE의 직접적인 영향을 측정하는 11개의 방법론을 비교하였다[20]. 그 중 대표적인 알고리즘은 다음과 같다[19]. 첫째, 특정 영역에 대해서 암환자들 중 CNAs가 있는 샘플들과 CNAs가 없는 샘플들로 나눈다. 그 후 CNAs가 있는 샘플들의 해당 영역의 유전자 발현량의 평균값과 CNAs가 없는 샘플들의 해당 영역의 유전자 발현량의 평균값의 차이를 표준편차로 나눈 값을 구한다. 이 값이 통계적으로 유의한지를 측정한다. 본 방법을 교모세포종 뇌암 환자들에게 적용했을 때, CNAs의 증가가 뚜렷한 영역의 경우 절반에 가까운 영역에서 유전자들의 발현

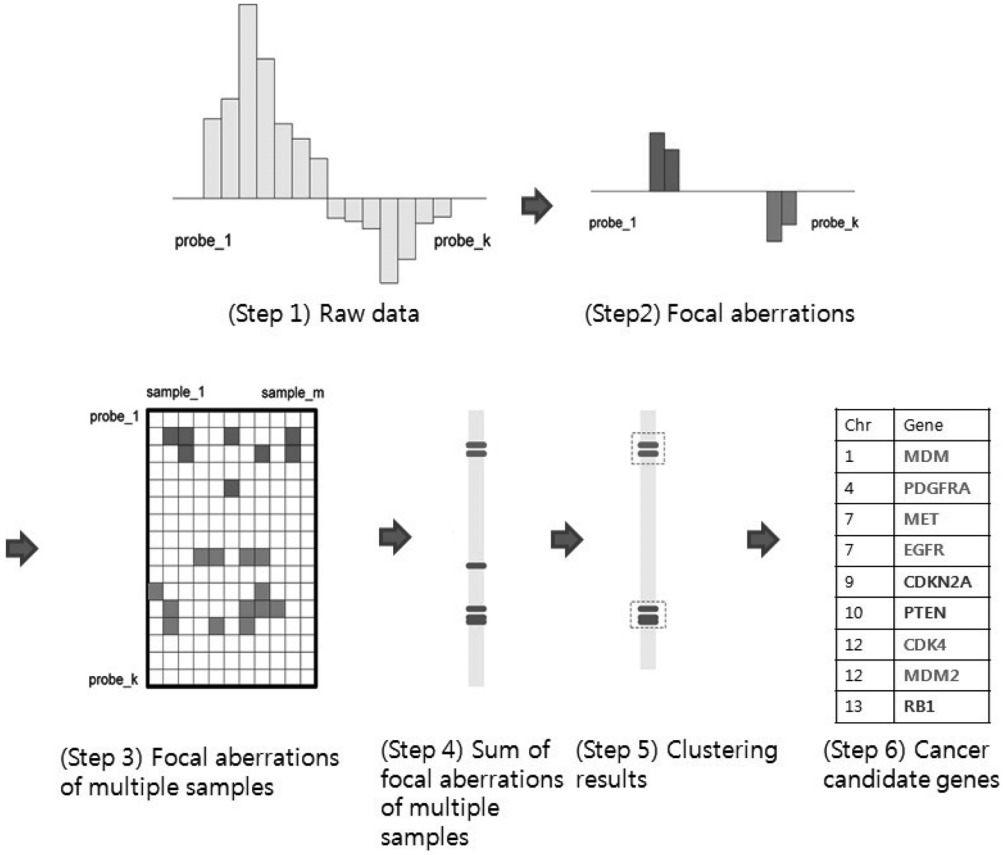

그림 6 WIFA 알고리즘(논문 [18]의 그림의 일부를 수정하였음)

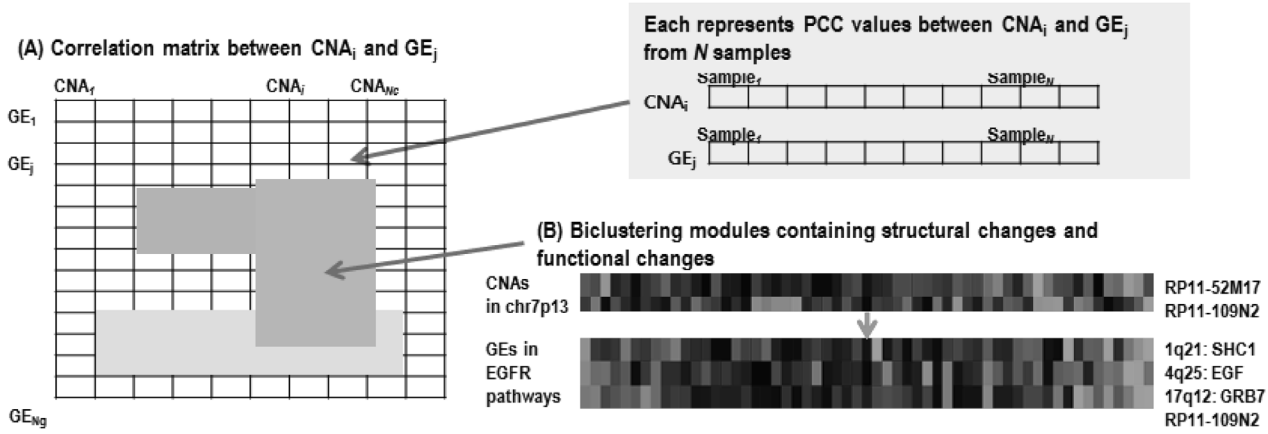

그림 7 바이클러스터링 방법론을 이용한 암세포에서의 CNAs와 GE사이의 관련성 연구
(논문 [22]의 그림의 일부를 수정하였음)

량이 증가하였다.

CNAs는 해당 유전자의 발현량에도 영향을 미칠 뿐만 아니라, 염색체의 다른 위치에 있는 상호 연관성이 높은 유전자의 발현량에도 영향을 미친다. 따라서 DNA의 구조적 변화가 유전자의 기능적 변화에 미치는 영향에 관한 연구들이 수행되고 있다[21,22]. 예를 들어, 흑색종 피부암(Melanoma)에서 두 유전자, TBC1D16와 RAB27A의 DNA 개수가 변한 경우, 세포의 증식이나 Vesicular Tracking, Melanogenesis 등에 관련된 유전자들의 발현량이 변화하는 것이 관찰되었다[21]. 또한 바이클러스터링(Biclustering)을 이용한 연구 방법론도 개발되었다[22]. 이 방법론은 그림 7에서 설명하였다. 첫째, 다수의 암환자로부터 구해진 CNAs 데이터와 유전자 발현형 데이터에 대해서, 모든 CNAs 데이터와 유전자 발현형 쌍에 대해서 상관관계(Pearson Correlation Coefficient, PCC)를 계산하여, 상관관계 매트릭스(Correlation Matrix)를 구한다. 둘째, 이 상관관계 매트릭스에 바이클러스터링 기법을 적용하여, 상관관계 유사도가 높은 DNA 영역들과 유전자들의 집합(Set)을 생성한다. 이 집합들을 모듈(Module)이라고 부른다. 바이클러스터링은 유사도가 높은 객체들을 특정 개수의 군집으로 클러스터링하는 기법의 변형된 기법으로써, 객체들이 다수의 클러스터에 속하는 것을 허용한다. 대부분의 유전자들은 다수의 기능을 가지고 다수의 생물학적 경로에 속하기 때문에, 바이클러스터링을 사용하여 모듈을 생성하였다. 각 모듈에 속한 유전자들은 CNAs와 유전자 발현형의 변화를 나타낸다. 이 방법론을 교모세포종 뇌암 환자에 적용하여 만든 모듈 중에서 하나는 chr7p13영역에서의 DNA 변화가 EGFR 생체신호경로의 변화와 연관이 있음을 보여주었다.

2.7 암 연구를 위한 CNAs, GE, 단백질 상호작용 데이터 통합 방법론

앞 장에서 설명한 CNAs와 GE에 단백질 상호작용 데이터를 통합하면 암과 관련된 생물학적 경로를 찾기 위한 모듈을 더 효율적으로 구성할 수 있다. 유전자들의 물리적 상호작용(Physical Interaction) 데이터를 네트워크 구조로 나타냈을 때, 많은 유전자들과 상호작용하는 허브 유전자(Hub Gene)가 암의 증식과 관련성이 높기 때문이다. GE, CNAs, PPI를 통합한 대표적인 연구는 다음과 같다[23]. 첫째, 각 유전자에 대해서 이 유전자를 대표하는 유전자를 선별하였는데, 이는 CNAs와 GE 데이터에서 PCC와 단백질 상호작용 네트워크에서 거리에 기반을 두었다. 둘째, 다수의 유전자들로부터 대표 유전자로 선발된 유전자들을 중심으로 모듈을 생성하였다. 그 후, 각 모듈 쌍에 대해서, 모듈들에 속한 유전자들의 PCC가 높고 단백질 상호작용 네트워크에서의 거리가 가까울 때, 모듈들을 통합하였다. 교모세포종 뇌암 및 난소암(Ovarian Cancer)에 본 방법론을 적용하였을 때, 각각 22개 23개의 모듈들이 생성되었고, 모든 모듈들이 생물학적 경로들과 연관성이 있었다.

3. 암 연구를 위한 텍스트 마이닝

3.1 생물학 텍스트 데이터의 축적과 텍스트 마이닝 기술의 응용

인터넷의 발달과 더불어 수많은 의학 및 생물학 분야의 문헌들을 온라인 공용데이터베이스에서 열람할 수 있게 되었고, 최근의 암 연구에 있어 텍스트 데이터베이스들은 필수적인 자원이 되었다. 생의학 문헌 데이터베이스인 PubMed(http://www.ncbi.nlm.nih.gov/

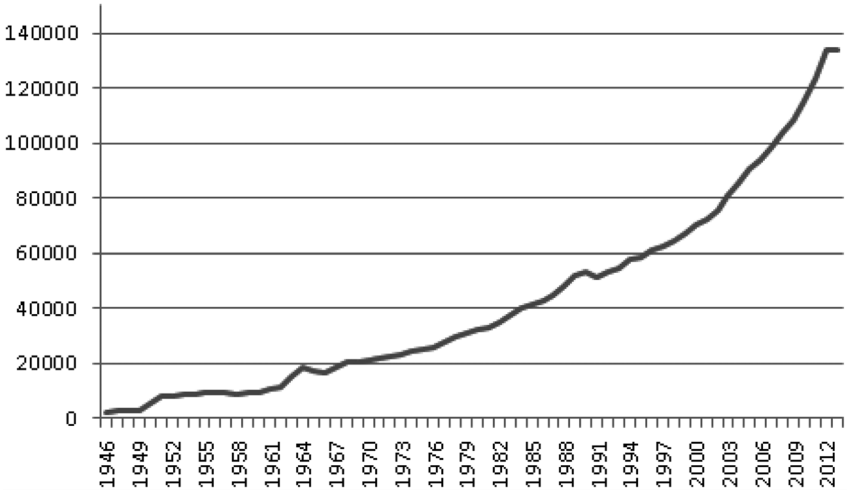

그림 8 PubMed에서 Cancer로 검색된 문헌의 연간 등록 수

pubmed)에서 Cancer를 키워드로 검색하였을 때 2014년 현재 약 290만 건의 암 관련 문헌이 존재한다. 그림 8에서 보는 것처럼 2008년 이후에는 매년 10만 건 이상 발표되고 시간이 지날수록 급격하게 늘고 있다. 이러한 방대한 텍스트 데이터로부터 유용한 정보를 추출하는 텍스트 마이닝 기술의 필요성이 점점 대두되고 있다[24]. 의생물학 문헌의 분석을 위해 사용되는 개체명 인식, 관계추출, 정보 검색 등의 텍스트 마이닝 기술들은 문헌을 분석하는데 드는 시간과 비용을 줄이는데 큰 기여를 하고 있다.

3.2 개체명 인식(Named Entity Recognition)

개체명(Named Entity)이란 문서에서 나타나는 고유한 의미를 가지는 명사나 숫자와 같은 고유한 성질의 표현을 뜻한다. 개체명 인식은 문서에서 개체명을 추출하고 추출한 개체명의 종류를 결정하는 것을 의미한다[25]. 생물학 문헌에서 개체명 인식 기법은 주로 유전자나 질병, 화학물질 등 생물학적 개체명을 찾는 것에 적용된다(그림 9). 생물학 문헌에서는 생물학적 개체명이 중심이 되어 연구의 중요 내용을 기술하기 때문에, 개체명을 인식하는 것이 우선적으로 수행되어야 한다.

가장 단순한 방식의 개체명 인식은 단어 사전을 구축한 후 사전 내의 단어들을 문헌에서 매칭 하는 방식이지만, 이는 사전에 없는 개체명은 인식할 수 없는 문제가 있다. 이런 문제를 해결하기 위한 방법으로 미국 국립 보건원(National Institutes of Health)에서 제공하는 Unified Medical Language System(UMLS)의 생물학 어휘들을 이용하는 방법이 있다. UMLS MetaMap Transfer[26]는 바이오메디컬 텍스트의 문장 구조를 분석한 후 개체명의 후보가 되는 단어와 UMLS의 단어를 매칭 한다. 단어사전의 문제를 해결하기 위해 기계학습을 이용한 알고리즘으로는 텍스트로부터 유전자와 단백질명을 인식하는 ABNER(A Biomedical Named Entity Recognizer)[27]와 BANNER[28]가 있다. 이 두 알고리즘은 통계적 모델인 조건부 랜덤 필드(Conditional Random Field, CRF)를 사용하는데 이는 확률 그래프 모델(Probabilistic Graphical Model)의 일종으로 텍스트 문장이나 염기서열 같은 순차데이터(Sequencial

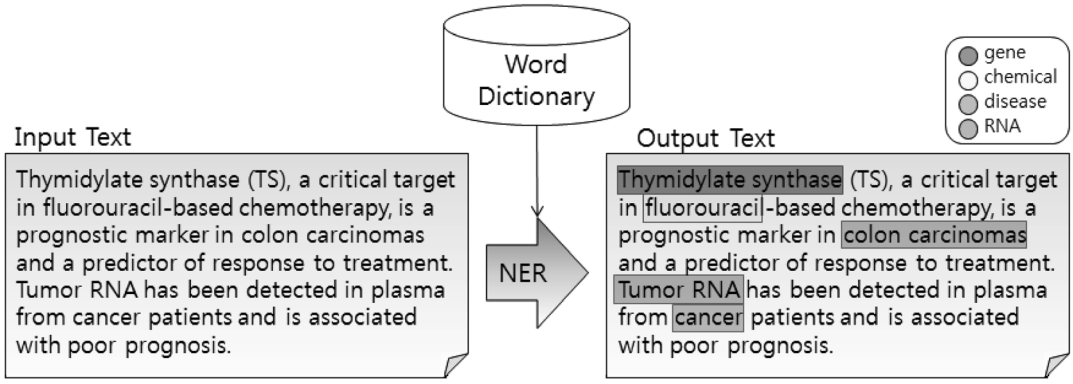

그림 9 단어사전을 이용한 개체명 인식의 예

Data)를 레이블링 하는데 사용된다. 단어사전과 기계학습 기법을 복합적으로 적용한 방법으로는 ChemSpot [29]이 있는데 이는 화학물질명을 인식하는 알고리즘으로 단어사전과 조건부 랜덤 필드를 같이 적용하여 정확도를 높이고자 하였다.

3.3 생물학적 개체명 표준화(Normalization)

생물학적 물질은 하나 이상의 개체명을 가질 수 있다. 다시 말하면 동일한 물질로라도 다른 텍스트에서는 약어로 쓰일 수도 있고 또는 전혀 다른 명칭이 사용될 수도 있다. 한 예로, 남성 호르몬 수용체(Androgen Receptor)라는 유전자는 AR이라는 약어로 사용될 수도 있고 DHTR, SBMA, NR3C4, SMAX1, Dihydrotestosterone Receptor와 같이 전혀 다른 개체명이 사용되기도 한다. 이와 같이 같은 물질이라도 개체명이 다를 때 이를 처리하지 않고 다른 물질로 취급하게 되면 추후의 분석에서 잘못된 결과를 초래한다. 개체명 표준화는 이런 문제를 해결하기 위한 처리기법이다.

개체명을 표준화하는 간단한 방법은 동의어 사전을 이용한 표준화이다. 이는 여러 데이터베이스의 동의어 및 약어와 비교하여 표준화를 수행하는데 간편하다는 이점이 있으나 사전을 이용한 개체명 인식에서와 같은 한계점이 있다. 이런 문제를 해결하기 위해 개체명이 있는 문장 또는 문서를 바탕으로 기계학습과 규칙 기반 기법(Rule-based Method)을 사용하는 연구들이 발표되었다.

유전자와 단백질에 대한 표현은 명칭이 같더라도 종에 따라 다른 물질을 가리키는 경우가 있어 표준화가 어렵다. 이를 해결하기 위해 다음과 같은 알고리즘들이 연구되고 있다. GeneTUkit[30]에서는 추출한 개체명에 대해 단어사전과 비교한 결과와 문서 내에서 같은 개체명의 전체 명칭(Full Name), 생물의 종에 대한 표현을 확인하여 표준화된 개체명을 도출한다. GNAT [31]은 표준화의 정확도를 높이기 위해 개체명을 문자별로 쪼개어 표준화된 개체명 후보를 찾고 잘못된 개체명이 될 수 있는 명칭들을 필터링하는 방법을 적용한다. Moara[32]는 생물 종 별로 데이터를 구분한 후 유전자와 단백질명을 표준화하여 정확도를 높이고, 단어사전 이용과 머신 러닝 기법을 이용한 방법을 병행하여 비교할 수 있도록 하였다.

3.4 관계 추출(Relationship Extraction)

관계 추출은 텍스트에서 개체명들 사이의 의미 있는 관계를 찾아내는 작업이다. 수많은 의학 및 생물학 문헌에서 유전자가 암에 미치는 영향 또는 유전자가 다른 유전자를 어떻게 조절하는지 등을 기술하고 있다. 그러나, 사람이 수십만 건에 달하는 방대한 문헌을 읽어서 정보를 파악하는 것은 막대한 시간과 비용이 소모되므로, 관계 추출기법을 이용하여 효율적으로 다양한 생물학적 요소들 사이의 관계를 파악하는 방법들이 개발되었다[33,34].

유전자와 질병간의 관계를 찾고자 하는 연구로 Al-Mubaid와 Singh의 연구[33]가 있는데, 대량의 문헌에서 질병과 유전자에 대한 개체명을 찾고, 발견된 개체명들이 함께 언급되는 빈도(Co-occurrence)를 기반으로 유전자와 질병간의 관계를 예측한다. 이 알고리즘은 질병과 유전자가 함께 언급될 때 두 개체간의 관계가 있을 것이라는 가정을 바탕으로 한다.

Turku Event Extraction System(TEES)[34]는 유전자와 유전자의 상태 변화(Biological Event)의 관계를 텍스트에서 찾는 프로그램이다. 유전자는 발현(Gene Expression), 조절(Regulation) 또는 인산화(Phosphorylation)와 같은 상태 변화를 통해 암에 영향을 미치게 되므로 암을 이해하기 위해 유전자의 상태 변화를 파악하는 것은 매우 중요하다. TEES는 이를 위해 유전자에 대한 개체명과 유전자의 상태 변화에 대한 개체명을 인식하고 유전자와 그 상태 변화의 관계를 문장 구조의 분석과 지지 벡터 머신(Support Vector Machine)을 이용하여 판별한다.

3.5 생물학 문헌 검색 엔진

검색 엔진은 사용자가 원하는 정보를 수집하여 제공하는 시스템으로 검색어를 바탕으로 데이터의 내용과 연관성이 높은 문헌들을 검색 결과로 반환하여 이용자들의 편의를 높인다. 의학 생물학 분야의 연구에서도 검색엔진은 중요한 역할을 하는데 의학 생물학 분야의 문헌을 찾는 것은 키워드 검색만으로는 필요한 문헌을 얻는데 한계가 있다. 원하는 문서를 정확히 찾기 위해 텍스트 마이닝 기법을 적용한 연구들이 발표되었다[35-38].

MedEvi 검색 엔진[35]은 PubMed에서 두 개 이상의 생물학적 개념들의 관계에 대한 명시적인 표현을 검색하여 사용자에게 제공한다. 사용자가 관심을 가진 개념들을 하나의 검색 질의어(Query)로 입력하면 개념을 지칭하는 단어 표현들이 가까운 문장들을 검색하여 결과로 출력한다. Medie[36]는 Subject-Verb-Object의 형식으로 검색 쿼리를 구성하며, 세 개념간의 관계를 표현하는 논문의 문장들을 보여주는 검색 엔진이다. GoPubMed[37]는 온톨로지를 적용한 검색 엔진으로, PubMed에 키워드 기반 웹 검색 엔진으로 가져온 검색 결과를 온톨로지 기술을 적용하여 계층적으로 분류하여 탐색이

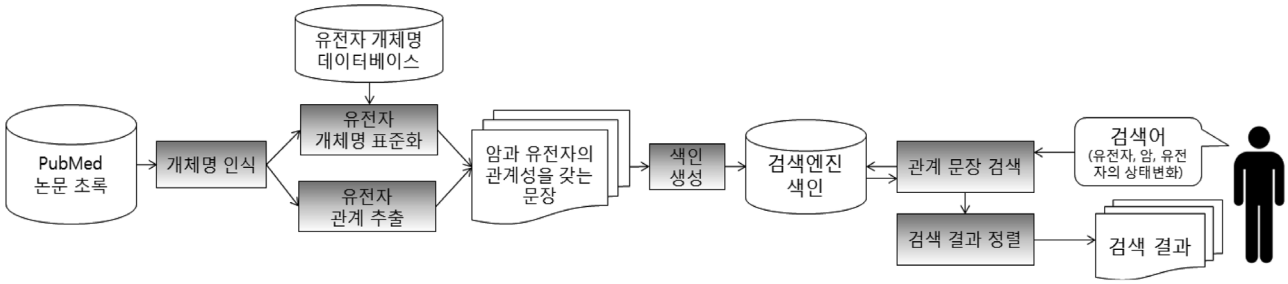

그림 10 DigSee의 구조도

용이하도록 만들었다.

DigSee[38]는 방대한 의생물학 문헌으로부터 암과 유전자의 관계를 나타내는 문장을 검색해주는 프로그램이다. 암에 관련된 유전자를 유전자의 상태 변화(유전자 발현형 변화, 유전자 조절 변화, 인산화 등)를 기반으로 하여 관련이 높은 문헌과 문장을 검색할 수 있다. DigSee의 전체적인 구조는 그림 10과 같다. 먼저 PubMed로부터 대량의 논문 초록들을 수집한 후 개체명 인식 기법과 관계 추출 기법을 사용하여 문헌들로부터 암과 유전자 사이의 관계를 나타내는 문장들을 추출한다. 암과 유전자 사이의 관계를 나타내는 문장들을 바탕으로 검색 엔진의 색인을 생성한다. 그 후 사용자가 검색어를 입력하면 색인을 바탕으로 관련 있는 문장들을 검색한 후, 암과 유전자의 관계가 잘 나타나는 문장 순으로 검색 결과를 정렬하여 사용자에게 반환한다.

DigSee에서는 사용자가 입력한 검색어와 암사이의 관련성을 판단하기 위해서, 베이지안 분류기(Bayesian Classifier)를 이용하였다. 분류기에서 사용된 특징(Feature) 들로는 개체명 인식의 정확도, 문장 구조에서 단어들 사이의 거리(유전자와 유전자의 상태 변화 사이의 거리, 상태 변화와 암 관련 단어 사이의 거리), 문장 내에 있는 다른 유전자와의 관계, 문장 내의 암의 특징을 표현하는 키워드들의 존재 여부, 문장이 연구의 목표나 실험 방법에 대한 내용을 포함하는지 여부 등의 열 가지 특징들이 있다.

예를 들면 “The expression of both HERV-E and ERV3 (another class of HERV) was detected in the same prostate carcinoma tissues”와 같은 문장은 그림 11에서처럼 트리로 문장 구조를 표현할 수 있다. 여기에서는 유전자와 유전자 상태 변화의 거리는 1이고 유전자 상태 변화와 암과의 거리는 3이며 찾고자 하는 유전자 외에도 다른 유전자가 문장 내에 있음을 볼 수 있다.

4. 고 찰

생물정보학 기술의 발전은 암의 연구에 커다란 기여를 하고 있다. 바이오테크놀로지의 발전으로 다양한 생물학 데이터가 축적되었고 이 데이터의 분석을 통해 암과 암에 영향을 미치는 유전자 사이의 관계가 밝혀지고 있다. 암과 연관성이 높은 유전자를 규명하기 위해 다양한 방법들이 적용되고 있는데 본 논문에서는 생물학 데이터를 분석하기 위한 방법들과 생물학 문헌을 분석하기 위한 방법들을 다루었다.

최근의 암 연구에는 NGS 데이터의 축적으로 시퀀싱

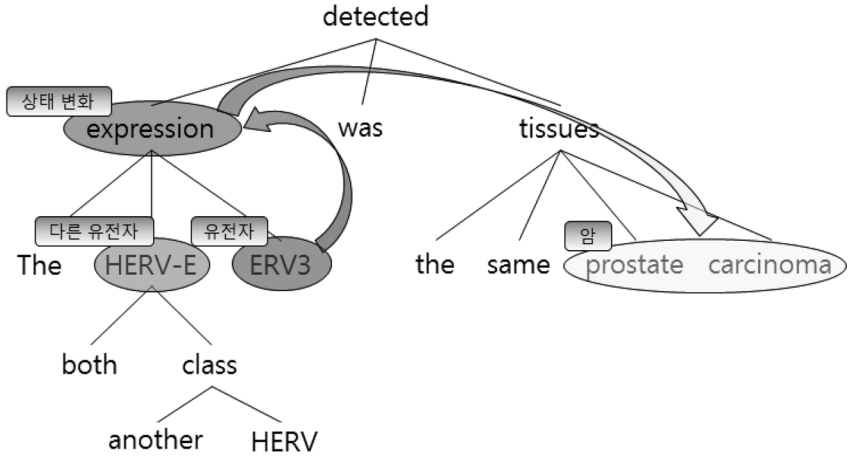

그림 11 예제 문장의 문장 구조 및 유전자와 유전자의 상태 변화, 암의 관계 표현

데이터를 이용한 데이터 마이닝 연구가 증가하고 있다. 시퀀싱 데이터를 이용할 경우 유전변이검출, 유전자의 발현량 측정, 후생유전적인 변형(Epigenetic Modification)의 측정을 바이오칩 보다 더욱 정확하게 할 수 있다는 장점이 있다. 그러나, NGS데이터는 데이터의 용량이 막대하여 이를 처리하기 위해서는 고용량의 저장 장치와 고성능의 컴퓨팅 자원이 필요하다.

또한, 생물학 문헌에 대한 텍스트 마이닝 연구는 전문(Fulltext)에 대한 접근이 어렵기 때문에 초록(Abstract) 데이터를 주로 이용하였다. 전문을 이용하는 것은 저작권 문제로 제한되는 경우가 많기 때문인데[39], PubMed Central[40]데이터 베이스와 같이 전문을 제공하는 경우가 늘어나고 있기 때문에 전문을 분석하여 유용한 정보를 추출하는 연구들이 점점 중요해지고 있다.

이와 같이 암 세포에서 추출된 분자 데이터와 생물학 문헌들에서 분석해야할 데이터의 크기가 점점 증가하고 있으므로, 방대한 생물학 데이터 및 문헌 데이터를 분석하기 위한 기술들은 앞으로 더 중요한 역할을 할 것이다.

참고문헌

[1] D. Sidransky, “Emerging molecular markers of cancer”, Nat Rev Cancer, Vol. 2, No. 3, pp. 210-219, 2002.

[2] A. H. Bild and et al., “Oncogenic pathway signatures in human cancers as a guide to targeted therapies”, Nature, Vol. 439, No. 7074, pp. 353-357, 2006.

[3] G. Ciriello and et al., “Emerging landscape of oncogenic signatures across human cancers”, Nat Genet, Vol. 45, No. 10, pp. 1127-1133, 2013.

[4] J. Barretina and et al., “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”, Nature, Vol. 483, No. 7391, pp. 603-607, 2012.

[5] J. Gao and et al., “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”, Sci Signal, Vol. 6, No. 269, pp. pl1, 2013.

[6] Tang, Ming-Xin, et al., “The APOE-∈4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics”, Jama 279.10: 751-755, 1998.

[7] Knuutila, Sakari, et al., “DNA copy number losses in human neoplasms”, The American journal of pathology 155.3: 683-694, 1999.

[8] Rovelet-Lecrux A, Hannequin D, Raux G, Meur NL, Laquerri?re A, et al., “APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy”, Nat Genet 38: 24-26, 2006.

[9] Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al., “Strong association of de novo copy number mutations with autism”, Science 316: 445-449, 2007.

[10] Fanciulli M, et al., “FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity”, Nat Genet 39: 721-723, 2007.

[11] Campbell PJ, Stephens PJ, Pleasance ED, O’Meara S, Li H, et al., “Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing”, Nat Genet 40: 722-729, 2008.

[12] Stone J, et al., “Rare chromosomal deletions and duplications increase risk of schizophrenia”, Nature 455: 237-241, 2008.

[13] Walters R, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11. Nature 463: 671-675, 2010.

[14] Olshen, Adam B., et al., “Circular binary segmentation for the analysis of array-based DNA copy number data”, Biostatistics 5.4: 557-572, 2004.

[15] Hupé, Philippe, et al., “Analysis of array CGH data: from signal ratio to gain and loss of DNA regions”, Bioinformatics 20.18: 3413-3422, 2004.

[16] Pawel Stankiewicz, James R. Lupski “Structural Variation in the Human Genome and its Role in Disease,” Annual Review of Medicine 61: 437-455, 2010.

[17] Beroukhim, Rameen, et al., “Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma”, Proceedings of the National Academy of Sciences 104.50: 20007-20012, 2007.

[18] Hur, Youngmi, and Hyunju Lee, “Wavelet-based identification of DNA focal genomic aberrations from single nucleotide polymorphism arrays”, BMC bioinformatics 12.1: 146, 2011.

[19] E. Hyman and et al., “Impact of DNA amplification on gene expression patterns in breast cancer”, Cancer Res, Vol. 62, No. 21, pp. 6240-6245, 2002.

[20] R. Louhimo and et al., “Comparative analysis of algorithms for integration of copy number and expression data”, Nat Methods, Vol. 9, No. 4, pp. 351-355, 2012.

[21] U. D. Akaviaand et al., “An integrated approach to uncover drivers of cancer”, Cell, Vol. 143, No. 6, pp. 1005-1017, 2010.

[22] H. Lee et al., “Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes”, Bioinforma-

tics, Vol. 24, No. 7, pp. 889-896, 2008.
[23] A. K. Azad and H. Lee, "Voting-based cancer module identification by combining topological and data-driven properties", PLoS One, Vol. 8, No. 8, pp. e70498, 2013.
[24] Cohen, Aaron M., and William R. Hersh., "A survey of current work in biomedical text mining", Briefings in bioinformatics 6.1, pp. 57-71, 2005.
[25] Chinchor, Nancy, et al., "1999 Named Entity Recognition Task Definition", MITRE and SAIC, 1999.
[26] Aronson, Alan R., "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program", Proceedings of the AMIA Symposium. American Medical Informatics Association, pp. 17, 2001.
[27] Settles, Burr, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text", Bioinformatics 21.14: pp. 3191-3192, 2005.
[28] Leaman, Robert, and Graciela Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition", Pacific Symposium on Biocomputing, Vol. 13, pp. 652-663, 2008.
[29] Rocktäschel, Tim, Michael Weidlich, and Ulf Leser, "ChemSpot: a hybrid system for chemical named entity recognition", Bioinformatics 28.12: pp. 1633-1640, 2012.
[30] Huang, Minlie, Jingchen Liu, and Xiaoyan Zhu, "GeneTUKit: a software for document-level gene normalization", Bioinformatics 27.7: pp. 1032-1033, 2011.
[31] Hakenberg, Jörg, et al., "Inter-species normalization of gene mentions with GNAT", Bioinformatics 24.16: pp. i126-i132, 2008.
[32] Neves, Mariana L., José-María Carazo, and Alberto Pascual-Montano, "Moara: a Java library for extracting and normalizing gene and protein mentions", BMC bioinformatics 11.1: pp. 157, 2010.
[33] Al-Mubaid, Hisham, and Rajit K. Singh, "A new text mining approach for finding protein-to-disease associations", American Journal of Biochemistry and Biotechnology 1.3: pp. 145, 2005.
[34] Björne, Jari, et al., "Extracting complex biological events with rich graph-based feature sets", Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, pp. 10-18, 2009.
[35] Kim, Jung-jae, Piotr Pezik, and Dietrich Rebholz-Schuhmann, "MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline", Bioinformatics 24.11: pp. 1410-1412, 2008.
[36] Miyao, Yusuke, et al., "Semantic retrieval for the accurate identification of relational concepts in massive textbases", Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1017-1024, 2006.
[37] Doms, Andreas, and Michael Schroeder, "GoPubMed: exploring PubMed with the gene ontology", Nucleic acids research 33.suppl 2: pp. W783-W786, 2005.
[38] Kim, Jeongkyun, et al., "DigSee: disease gene search engine with evidence sentences(version cancer)", Nucleic acids research 41.W1: pp. W510-W517, 2013.
[39] Hirschman, Lynette, et al., "Overview of BioCreAtIvE: critical assessment of information extraction for biology", BMC bioinformatics 6.Suppl 1: pp. S1, 2005.
[40] Beck, Jeff, "Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles", International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, Canada. 2010.

약 력

장 호

2010 전북대학교 컴퓨터공학과 졸업(학사)
2012 광주과학기술원 정보통신공학부(석사)
2012~ 현재 광주과학기술원 정보통신공학부 박사과정
관심분야 : 생물정보학, 유전변이 검출, 질병 예측
E-mail : walker83@gist.ac.kr

김 정 균

2011 한국항공대학교 컴퓨터정보공학 전공 졸업(학사)
2013 광주과학기술원 정보통신공학부 졸업(석사)
2013~현재 광주과학기술원 정보통신공학부 박사과정
관심분야 : 텍스트 마이닝, 정보 추출, 생물학적 자연어 언어 처리 기법
E-mail : jeongkyunkim@gist.ac.kr

이 현 주

1997 한국과학기술원 전산학과 졸업(학사)
1999 서울대학교 컴퓨터공학과 졸업(석사)
2006 미국 Univ. of Southern California 졸업(박사)
2007~현재 광주과학기술원 정보통신공학부 부교수
관심분야 : 생물정보학, 암 정보학, 기계학습
E-mail : hyunjulee@gist.ac.kr