문헌정보와 유전자 발현 및 상호 작용 데이터를 통합, 암의 단계를 고려한 질병 유전자 예측 방법

김정림, 여윤구, 박상현*
연세대학교 컴퓨터과학과
e-mail : kimgogo02@cs.yonsei.ac.kr

The gene prediction method considering stages of cancer, obtained by integrating gene expression, genetic interaction data and document

Jungrim Kim, Yunku Yeu, Sanghyun Park*
Dept. of Computer Science, Yonsei University

요 약

유전체에 대한 관심이 크게 증가하면서, 이에 따른 다양한 연구가 이루어졌다. 그 결과 유전체와 관련된 다양한 종류의 데이터가 얻어졌으며, 그것을 해석하고 다른 데이터와 통합하는 것이 중요한 연구과제 중 하나가 되었다. 본 논문은 유전자 상호작용(genetic interaction) 데이터, 유전자 발현 데이터, 문헌으로부터 텍스트마이닝 기술을 통해 얻은 이종(heterogeneous) 데이터를 통합하여 암과 관련이 있는 유전자를 찾는 실험을 수행하였다. 또한, 단순히 질병(disease)-정상(normal)의 대조가 아니라 암의 단계(stage)를 고려한 실험을 수행하였다. 데이터를 통합하지 않거나 암의 단계를 고려하지 않았을 경우에 비하여 제안하는 방법이 더 높은 유전자 예측 성능을 나타냈다.

1. 서론

게놈 프로젝트 (Genome project) 이후로 유전체에 대한 관심이 증가하면서 다양한 종류의 데이터가 등장하였고, 이러한 데이터를 해석하고, 다른 데이터와 통합하는 것은 중요한 연구과제이다. 또한, 다양한 데이터들을 통한 질병유전자 예측은 암과 같은 복잡한 유전적 질병의 진단, 치료, 예방, 신약재창출(Drug repositioning) 등에 핵심적인 역할을 수행 할 수 있기 때문에 지속적으로 연구되고 있는 분야이다.

유전체에 대한 정보는 다양한 데이터를 이용해서 측정할 수 있다. DNA 마이크로어레이[1] 기술은 수만 개의 유전자의 발현을 동시에 관찰 할 수 있는 수단이며 피어슨 상관계수(Pearson Correlation Coefficient) 및 발현 차이(differential expression) 등을 이용하여 유전자 간의 연관 관계를 추정할 수 있다. [2, 3] 또한, 유전자 상호작용 데이터는 생물체의 표현형 발현에 영향을 주는 유전자들간의 상호작용을 기록해놓은 데이터로 질병유전자 예측을 포함한 다양한 영역에서 사용되고 있다. 텍스트마이닝 기술은 유전체에 관한 방대한 양의 문헌정보로부터 중요한 의미가 있는 데이터를 새롭게 얻어 낼 수 있으며, 이러한 데이터들은 유전자 예측, 신약재창출 등에 활용되고 있다.

위에 언급된 데이터들은 모두 질병유전자 예측 연구에 사용될 수 있지만, 각각의 데이터들은 모두 단점을 가지고 있다. 특히, 마이크로어레이 데이터의 경우 노이즈가 매우 많이 포함되어 있기 때문에 이를 단독으로 사용할 경우에 정확도 높은 질병 유전자 예측을 하기 힘들다. 이러한 점을 보완하기 위하여 본 논문은 텍스트마이닝 기법을 통해 문헌정보로부터 얻은 데이터와 유전자들 간의 상호작용 데이터를 마이크로어레이 데이터와 함께 이용하여 일부 알려진 질병 유전자와 상호작용이 존재하는 유전자들에 한해서 점수를 매겨 질병 유전자를 예측하였다. 그 결과, 이들 데이터들을 단독으로 사용했을 때보다 정확도 높은 유전자 예측을 하였다. 또한, 마이크로어레이 상에서 암의 특정 단계에서만 발현되는 질병 유전자가 있다고 가정하고 암의 단계가 구분된 마이크로어레이 데이터를 이용하여 정확도 높은 실험 결과를 얻었다.

* : 교신저자, e-mail: sanghyun@cs.yonsei.ac.kr
※ 이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2012R1A2A1A01010775).

2. 데이터 출처 및 실험 방법

2.1 데이터 출처

여러 생물학 데이터를 통합하여 유전자 예측에 이용하기 위하여 아래와 같은 데이터를 수집하였다.

BioGRID[4]에서 제공하는 BioGRID Release 3.2.104 버전의 유전자 상호작용 데이터를 다운로드 받았다. 다운로드 받은 데이터에서 사람의 유전자 상호작용 중 중복된 관계(relation)와 자기 자신과 관계가 있는 것을 제외한 13,993 개의 유전자로 이루어진 127,688 개의 유전자 상호작용 관계를 얻었다.

GEO (Gene Expression Omnibus)[5]에서 배포된 결장암 마이크로어레이 데이터 GSE21815[6]를 다운로드

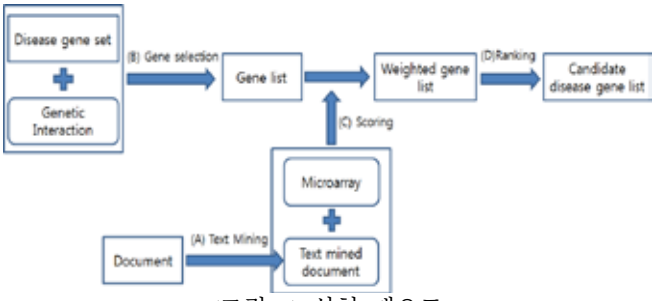
받았으며, 총 132 개의 샘플 중 샘플에 대한 정보가 명확한 정상 조직 샘플 9 개, 결장암 1 단계 조직 샘플 12 개, 2 단계 조직 샘플 27 개, 3 단계 조직 샘플 16 개, 4 단계 조직 샘플 11 개를 사용하였다. 또한, 마이크로어레이 데이터 상에서 유전자 기호(gene symbol)를 가지는 행(row)만을 실험에 사용하였고, 복수 개의 행이 동일한 유전자 기호를 나타낼 경우 해당 유전자 기호를 나타내는 행들의 값의 평균을 구하여 하나의 행으로 통합하였다.

HGNC (HUGO Gene Nomenclature Committee)[7]에서 배포된 41,321 개의 유전자 기호를 포함하는 유전자 정보 데이터를 다운로드 받았다.

PubMed[10]에서 결장암(colorectal cancer)의 키워드를 포함하고 있는 157,740 의 논문의 개요를 다운로드 받았다.

GeneCards version 3.10 [8,9]에서 결장암(colorectal cancer)의 검색 결과로 나온 유전자 3,272 개를 다운로드 받아서 그 중 2,024 개를 검증을 위해 사용하였다.

2.2 실험 방법



(그림 1) 실험 개요도

실험은 그림 1 과 같이 크게 (A) 문헌정보 처리(text mining), 질병 유전자 셋 (disease gene set)을 이용한 (B) 유전자 선택(gene selection), 트레이닝 셋(training set)을 통한 (C) 점수화(scoring), (D) 순위화(ranking)의 과정으로 구성되어 있다. 트레이닝 셋을 만들기 위해서 KEGG (Kyoto Encyclopedia of Genes and Genomes)[11]와 NCI(National Cancer Institution) [12]로부터 질병 유전자 셋을 얻었다. 다음으로 점수화 과정에서는 NCI 유전자 셋과 KEGG 유전자 셋 중 하나는 트레이닝 셋으로 사용하고, 다른 하나는 검증을 위해 사용하는 교차 타당화 (cross validation) 실험을 수행하였다.

2.2.1 문헌정보 처리

문헌 정보에서 질병과 관련된 정보를 추출하기 위하여, POS(Part of Speech) tagger[13]를 이용하여 논문에 등장하는 명사를 식별하였다. 식별된 명사를 HGNC 에서 제공하는 유전자 데이터베이스의 유전자 기호와 비교하여 각각의 유전자 출현 빈도수(frequency)를 포함하는 유전자 목록을 얻었다.

2.2.2 유전자 선택

KEGG 와 NCI 을 통해서 표 1 에 나와있는 결장암과 관련된 질병 유전자의 질병 유전자 셋을 얻었다. 다음으로 유전자 상호작용 데이터 상에서 표 1 에 나와있는 질병 유전자 셋과 상호작용이 존재하는 유전자들을 찾고 이후에 이들 유전자에 한해서 점수화(scoring)을 수행하였다.

<표 1> KEGG 와 NCI 에서 얻은 결장암과 높은 관련이 있는 유전자들의 질병 유전자 셋

KEGG	NCI
Oncogenes genes CTNNB1(β − catenin), KRAS **Tumor suppressor genes** APC, DCC, TGFBR2, SMAD2, SMAD4, BAX, TP53 **DNA repair genes** MLH1, MSH2, MSH3, MSH6	**Tumor suppressor genes** APC, TP53, NEK4(STK11), PTEN, BMPR1A, SMAD4 **Repair/stability genes** MYH(MUTYH), MLH1, MSH2, MSH6, PMS2, EPCAM

2.2.3 점수화

이 단계에서는 암의 단계와 질병-정상 상태에서의 차등 발현 정도, 문헌 데이터에서의 출현 빈도 등을 종합적으로 고려하기 위해 각 요소를 점수로 환산하였다. 먼저 유전자 발현 데이터 상에서 질병 유전자와 임계 값(threshold) 이상의 피어슨 상관계수를 갖는 유전자들을 선택하여 이들의 평균 상관계수인 Score_{PCC_i}를 계산하였다. 이 값은 유전자와 알려진 질병 유전자와의 연관 정도를 나타내며, 유전자가 항상 활성화되어 있지 않을 수 있으므로, 질병의 각 단계별 데이터를 구별하여 계산하였다. 다음으로 정상세포와 각 단계의 암세포에서 임계 값 이상으로 발현 차이를 갖는 유전자들을 선택하여 이들의 평균 상관계수 $\text{Score}_{\text{diff}_i}$를 계산하였다. $\text{Score}_{\text{diff}_i}$는 유전자와 질병 간의 연관 정도를 나타낸다. 질병 유전자와의 상관계수, 질병-정상간의 차등 발현 정도는 다음과 같이 구하였다.

$$\text{Score}_{PCC_i} = \frac{sum_{stage=1,2,34}\, P(i)}{N_{i,PCC}}$$

$$P(i) = \begin{cases} if\ PCC_{i,stage} > \delta_{PCC},\ P(i) = PCC_{i,stage} \\ else, P(i) = 0 \end{cases} \quad (1)$$

$$\text{Score}_{\text{diff}_i} = \frac{sum_{stage=1,2,34}\, D(i)}{N_{i,\text{diff}}}$$

$$D(i) = \begin{cases} if\ \text{diff}_{i,stage} > \delta_{\text{diff}},\ D(i) = \text{diff}_{i,stage} \\ else, D(i) = 0 \end{cases} \quad (2)$$

score_{PCC_i} : 유전자 i 의 피어슨 상관계수 점수
$\text{score}_{\text{diff}_i}$: 유전자 i 의 발현 차이 점수
$\text{PCC}_{\text{i,stage}}$: stage 에서 유전자 i 의 피어슨 상관계수
$\text{diff}_{\text{i,stage}}$: stage 에서 유전자 i 의 발현 차이 값
δ_{PCC} : 피어슨 상관 계수의 임계 값
δ_{diff} : 발현 차이 값의 임계 값
$N_{i,PCC}$: $P(i)$ 의 값이 0 이 아닌 stage 의 개수
$N_{i,\text{diff}}$: $D(i)$ 의 값이 0 이 아닌 stage 의 개수

암의 단계를 고려하지 않거나 단계가 구분되어 있지 않는 마이크로어레이 데이터의 경우 점수화 과정에서 정상, 암 1 단계, 2 단계, 3 단계, 4 단계로 구분하여 점수화하지 않고, 단순히, 암과 정상으로 구분하여 점수화를 하였다.

마지막으로 문헌정보로부터 얻은 유전자들의 출현 빈도 값은 편차를 보정하기 위해서 다음과 같은 처리를 하였다. Score_{freq_i}는 질병와 유전자 간의 연관 관계를 의미한다.

$$\text{Score}_{freq_i} = \log_{10}(freq_i + 1) \quad (3)$$

$\text{score}_{\text{freq}_i}$: 유전자 i 의 출현 빈도 점수
freq_i : 유전자 i 의 출현 빈도

위에서 구한 Score_{PCC_i}와 $\text{Score}_{\text{diff}_i}$, Score_{freq_i}를 각각 최대 최소 정규화 (min-max normalization) 처리한 후, 각각 가중치를 다르게 주어 다음과 같이 각 유전자의 점수를 계산하였다.

$$\text{Score}_i = w_1 * Score_{freq_i} + w_2 * Score_{PCC_i} + w_3 * Score_{\text{diff}_i} \quad (4)$$

score_i = *유전자 i의 점수*
w_1 : 출현 빈도 점수의 가중치
w_2 : 피어슨 상관 계수 점수의 가중치
w_3 : 발현 차이 점수의 가중치

2.2.4 순위화

최종적으로, 유전자 상호작용 데이터를 통해 얻은 유전자들을 점수에 따라서 순위화를 하였다.

3. 실험 결과

순위화를 위한 가중치는 $w_1 = 1, w_2 = 1, w_3 = 2$ 을 사용하였으며 임계 값은 $\delta_{PCC} = 0.5$, $\delta_{\text{diff}} = 0.2$ 을 사용하였다.

먼저 데이터 통합 방법의 효과를 입증하기 위해서 다음과 같은 실험을 수행하였다. NCI 와 KEGG 의 질병 유전자 셋 중 하나를 트레이닝 셋으로 사용하여 질병 유전자를 예측한 다음, 그 결과 중에 다른 유전자 셋이 얼마나 포함되어 있는지를 확인하였다.

NCI 유전자 셋을 이용하여 실험을 수행하였을 때, 전체 질병 유전자 셋에서 NCI 유전자를 제외한 질병 유전자 7 개를 찾는 실험에 대해서 다음과 같은 결과를 얻었다.

<표 2> NCI 질병 유전자 셋을 이용하였을 때의 결과

	Microarray	Document	Microarray + GI + Document
Top 30	1	2	2
Top 50	1	3	5
Top 100	1	3	7

표 2 에서 보이는 바와 같이 마이크로어레이 또는 문헌 정보만을 단독으로 이용했을 때와 비교하여서 점수가 높은 상위 30 개의 유전자를 뽑았을 때를 제외하고, 데이터를 통합하여 사용했을 때 더 많은 수의 질병 유전자를 찾았다. 특히, 상위 100 개의 유전자를 뽑았을 때 질병 유전자 7 개를 모두 찾았다.

반대로 KEGG 유전자 셋을 이용하여 동일한 방법으로 전체 질병 유전자 셋에서 KEGG 유전자를 제외한 질병 유전자 6 개를 찾는 실험에 대해서 다음과 같은 결과를 얻었다.

<표 3> KEGG 질병 유전자 셋을 이용하였을 때의 결과

	Microarray	Document	Microarray + GI + Document
Top 30	0	0	3
Top 50	0	1	4
Top 100	0	3	5

표 3 에서 보이는 바와 같이 데이터를 통합하여 상위 30 개, 50 개, 100 개의 유전자를 뽑았을 때 모두 더 많은 수의 질병 유전자를 찾았다.

다음으로 암의 단계를 고려하였을 때의 예측 성능 향상 정도를 입증하기 위하여, 암의 단계를 고려하여 $\text{Score}_{\text{PCC}_i}$와 $\text{Score}_{\text{diff}_i}$ 을 계산했을 때와 고려하지 않고 계산했을 때의 성능을 비교하는 실험을 하였다. GeneCards 을 통해서 얻은 결장암과 관련이 있는 유전자가 실험 결과에 얼마나 많이 포함되어 있는지 확인하는 실험을 하였을 때, 다음과 같은 결과를 얻었다. 표 4 에서 보이는 바와 같이 상위 10 개, 30 개, 50 개, 100 개의 유전자를 뽑았을 때 모두 암의 단계를 구분하여 실험하는 경우 더 많은 질병 유전자를 찾았다.

<표 4> 스테이지를 구분하여 Score_{PCC_i}와 $\text{Score}_{\text{diff}_i}$을 계산했을 때와 구분하지 않고 계산했을 때의 성능 비교 결과

	단계를 구분하였을 때	단계를 구분하지 않을 때
Top 10	9	6
Top 30	26	23
Top 50	40	35
Top 100	70	59

4. 결론

본 논문이 제안하는 방법은 유전체에 관해서 다양한 연구를 통해 얻어진 데이터를 통합하여 보다 정확한 질병 유전자 예측을 하는 것이다. 실험 결과, 마이크로어레이 또는 문헌자료를 단독으로 사용하였을 때보다 데이터를 통합하여 사용했을 때에 더욱 정확도 높은 질병 유전자 예측이 가능하였다. 또한, 실험 과정에서 암의 단계를 고려하여 유전자의 관련 정도를 계산했을 때, 더욱 정확도 높은 실험결과를 얻을 수 있었다. 이번 연구에서는 다양한 데이터들을 통합하여 사용하는 방법에 중점을 두었다면, 차후 연구에서는 각각의 데이터를 효과적으로 통합하기 위한 전처리(preprocess) 방법과 암의 단계에 따라서 다르게 발현되는 유전자들을 효과적으로 해석하는 방법을 연구 할 계획이다.

참고문헌

[1] Duggan. D. J, Bittner. M, Chen Y, Meltzer. P, Trent. J. M,
"Expression profiling using cDNA microarrays," Nature Genetics Supplement, vol. 21, pp.10-14, 1999.
[2] Eunji Shin, Yongmi Yoon, Jaegyoon Ahn, Sanghyun Park, "TC-VGC: A Tumor Classification System using variations in Genes' Correlation",Comput. Methods Programs in Biomed, vol. 104, pp.87-101, 2011
[3] Liu X, Liu ZP, Zhao XM, et al "Identifying disease genes and module biomarkers by differential interactions" J Am Med Inform Assoc, vol. 19, pp.241-248,2012
[4] Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust JM, Livstone MS, Oughtred R, Dolinski K, Tyers M. **"**The BioGRID Interaction Database: 2013 update" Nucleic Acids Res. 2012 Nov 30
[5] Edgar R, Domrachev M, Lash AE. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", Nucleic Acids Res. 2002 Jan 1;30(1):207-10
[6] "Mori M, Mimori K, Yokobori T et al., 2011, "Gene expression profiles in 132 laser microdissected colorectal cancer tissues" (data accessible at NCBI GEO database (Edgar et al., 2002), accession GSE21815)."
[7] HUGO Gene Nomenclature Committee at the European Bioinformatics Institute, http://www.genenames.org/cgi-bin/hgnc_downloads/
[8] GeneCards http://www.genecards.org/
[9] Belinky, F, Bahir, I, Stelzer, G, Zimmerman, S, Rosen, N, Nativ, N, Dalah, I, Iny Stein, T, Rappaport, N, Mituyama, M, Safran, M and Lancet, D. "Non-redundant compendium of human ncRNA genes in GeneCards", Bioinformatics 29, 2: 255-61 (2013)
[10] PubMed, http://www.ncbi.nlm.nih.gov/pubmed
[11] KEGG, http://www.kegg.jp/kegg/kegg2.html
[12] National Cancer Institute, "Genetics of Colorectal Cancer", http://www.cancer.gov/cancertopics/pdq/genetics/colorectal/HealthProfessional/page2, 2013. 7. 25
[13] Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/