RNA 시퀀스 데이터와 단백질 상호관계 데이터를 활용한 체세포 돌연변이 데이터 보완 방법

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

Method for compensating somatic mutation data using RNA-sequence data and protein interaction data

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

요 약

유전체에 대한 관심이 증가함에 따라서 방대한 양의 사용 가능한 생물학적 데이터가 등장하고 있으며, 이를 활용하는 다양한 연구가 진행되고 있다. 특히, 최근에 등장한 체세포 돌연변이 데이터는, 체세포 상에서 발생한 missense, nonsense, slience 돌연변이 등을 기록해놓은 데이터로, 이는 질병을 발생시키는 causal event를 연구하는데 활용 될 수 있다. 그렇지만, 이는 특정 질병을 앓고 있는 환자 중 다수의 체세포 돌연변이를 갖는 환자는 매우 적으며, 환자 간 공통적으로 가지고 있는 체세포 돌연변이 또한 매우 적은 한계를 가지고 있다. 본 논문에서는 단백질 상호작용 데이터와 각각의 체세포 돌연변이를 가지고 있는 환자의 RNA 시퀀스 데이터를 활용하여 체세포 돌연변이 유전자 관계를 찾음으로써 이러한 한계점을 극복하고자 노력하였다.

1. 서 론

사용 가능한 방대한 양의 생물학적 데이터가 등장함에 따라서, 이를 활용하여 생물학적으로 가치 있는 새로운 결과를 찾는 많은 연구들이 있었다. 특히, 질병을 발생시키는 유전자 모듈을 찾는 연구는 질병을 정복하는데 중요한 역할을 하기 때문에 많은 연구들이 있었다. 그렇지만, 암과 같은 질병은 단순히 한가지 유전자 모듈이 암의 전체 발생과정에 영향을 주지 않기 때문에 이를 연구하는 것은 많은 어려움이 있다. [1] 최근에 등장한 체세포 돌연변이 데이터는, 체세포 상에서 발생한 missense, nonsense, slience 돌연변이 등을 기록해놓은 데이터로, 특정 유전자 영역에 돌연변이가 발생하면 '1' 그렇지 않으면 '0'의 digital signal형태로 표현 가능하며, 체세포 상의 실질적인 돌연변이를 기록해놓은 데이터이기 때문에 질병을 발생시키는 causal event로 간주할 수 있다. 그렇지만, 특정 질병을 앓고 있는 환자 중 다수의 체세포 돌연변이를 갖는 환자는 매우 적으며, 환자 간 공통적으로 가지고 있는 체세포 돌연변이 또한 매우 적은 한계를 가지고 있다. Matan Hofree [2] 은 이러한 한계점을 극복하기 위하여 유전자 상호작용 데이터를 이용하여 유전자 네트워크를 만들고, 각각의 환자의 체세포 돌연변이가 발생한 영역의 유전자를 시드(Seed)로 사용하여 smoothing기법을 적용하여 해결하였다. 본 논문에서는 단백질 상호작용 데이터와 RNA 시퀀스 데이터를 활용하여 체세포 돌연변이 유전자 관계를 찾고, 해당 체세포 돌연변이 유전자 관계를 기존 체세포 돌연변이 데이터와 함께 사용하여 체세포 돌연변이 데이터가 가지는 한계를 보완하는 방법을 제안한다.

* : 교신저자, e-mail: sanghyun@yonsei.ac.kr
※ 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015R1A2A1A05001845).

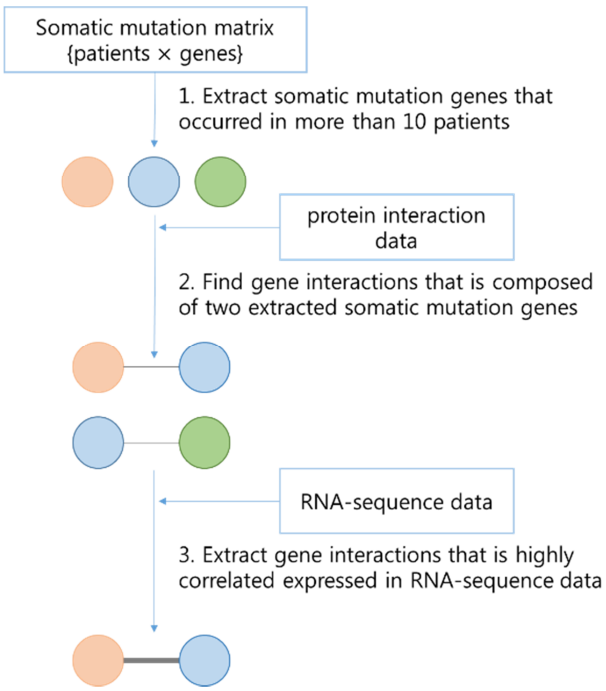
2. 데이터 전처리

본 논문에서는 크게 체세포 돌연변이 데이터, RNA 시퀀스 데이터, 단백질 상호작용 데이터를 사용하였다. 먼저, TCGA (The cancer genome atlas) [3]에서 결장암, 유방암, 폐암의 체세포 돌연변이 데이터와 RNA시퀀스를 다운받아서 사용하였으며, 단백질 상호작용 데이터의 경우 STRING [4] 에서 version. 10을 다운 받아 사용하였다. 체세포 돌연변이 데이터는 제안하는 방법론에 적용하기 위하여 행으로 체세포 돌연변이가 발생한 영역의 유전자, 열로 각각의 환자로 구성된 행렬의 형태로 변환을 하였으며 환자가 특정 체세포 돌연변이를 가지고 있으면 '1', 그렇지 않으면 '0'의 값을 주었다. RNA 시퀀스 데이터의 경우 제안하는 방법론에 적용하기 위하여 행으로 유전자, 열로 각각의 환자로 구성된 행렬의 형태로 변환을 하였다. 마지막으로, 단백질 상호 작용데이터의 경우 사람의 단백질 상호 작용 데이터 중 self-interaction을

제거하여 사용하였다.

3. 방법

실험은 그림 1과 같이 크게 (1) 체세포 돌연변이유전자 추출, (2) 체세포 돌연변이 유전자 관계 탐색, (3) 유의미한 체세포 돌연변이 유전자 관계 추출로 구성되어 있다.



(그림 1) 실험 개요도

3.1 체세포 돌연변이 유전자 추출

체세포 돌연변이 행렬로부터 10명 이상의 환자에게서 공통적으로 발견되는 체세포 돌연변이 유전자를 추출한다.

3.2 체세포 돌연변이 유전자 관계 추출

단백질 상호작용 데이터로부터 앞의 단계에서 추출한 체세포 돌연변이 유전자들로 이루어진 상호작용 관계를 찾는다.

3.3 유의미한 체세포 돌연변이 유전자 관계 추출

본 단계는 이전 단계에서 찾은 체세포 돌연변이 유전자 관계들 중 유의미한 관계를 추려내는 단계이다. 이를 위해서 이전 단계에서 찾은 두 체세포 돌연변이 유전자들이 RNA 시퀀스 데이터 상에서 얼마나 유의미하게 상호작용하는지 확인하였으며, 이를 위해서 피어슨 상관계수를 이용하였다. 피어슨 상관계수는 두 변수간의 선형적 관계를 분석하는데 보편적으로 사용되는 통계적 분석 방법으로, 두 변수가 양의 상관관계를 가지면, 1까지의 양의 값을 가지고, 음의 상관관계를 가지면 -1까지의 음의 값을 가진다. 이때 상관계수의 절대값이 1에 가까워 질수록 강한 상관관계를 가지는 것을 의미한다. 본 방법에서는 피어슨 상관계수의 절대값이 0.3 이상인 체세포 돌연변이 유전자 관계를 추출하였으며, 피어슨 상관계수를 구하는 과정에서 전체 환자 샘플이 아닌 상관계수를 구하는 체세포 돌연변이를 실질적으로 가지고 있는 환자의 샘플만을 사용하여 측정하였다.

4. 실험 결과

실험 결과로 3.1에서 찾은 결장암의 체세포 돌연변이 유전자 2472개로부터 1000개, 폐암 체세포 돌연변이 유전자 5416개로부터 2020개, 유방암 체세포 돌연변이 유전자 1623개로부터 584개의 새로운 체세포 돌연변이 관계를 찾을 수 있었다.

5. 결론

체세포 돌연변이 데이터는 질병을 발생시키는 causal event를 연구하는데 유용하게 활용될 수 있지만, 특정 질병을 앓고 있는 환자 중 다수의 체세포 돌연변이를 갖는 환자가 매우 적다는 한계와 환자 간 공통적으로 공유하고 있는 체세포 돌연변이 수가 매우 적다는 한계를 가지고 있다. 본 논문에서는 이러한 한계점을 보완하기 위하여 단백질 상호관계 데이터와 RNA 시퀀스 데이터를 활용하여, 체세포 돌연변이 유전자 관계를 찾음으로써 체세포 돌연변이 데이터가 가지는 한계를 보완하고자 하였다. 즉, AXIN1 유전자 영역에 체세포 돌연변이를 가지고 있는 환자 A와 CTNNB1 유전자 영역에 체세포 돌연변이를 가지고 있는 환자 B가 있다면, 기존 데이터로는 두 환자간의 유사성을 찾기 힘들었지만, AXIN1-CTNNB1 등과 같은 체세포 돌연변이 관계를 찾음으로써 두 환자 사이의 유사성을 찾을 수 있다. 본 논문에서 찾은 관계는 체세포 돌연변이 데이터를 활용한 환자의 clustering 또는 classification 등에 사용 될 수 있다. 또한, 단백질 상호작용 데이터뿐만 아니라 문헌정보 등을 추가적으로 사용한다면, 더 나은 체세포 돌연변이 유전자 관계를 찾을 수 있을 것이다.

참고문헌

[1] Toshinori Hinoue, Daniel J. Weisenberger, Christopher P.E. Lange, Hui Shen, Hyang-Min Byun, David Van Den Berg, Simeen Malik, Fei Pan,Houtan Noushmehr, Cornelis M. van Dijk, Rob A.E.M. Tollenaar, and Peter W. Laird, *Genome-scale analysis of aberrant DNA methylation in colorectal cancer*, Genome Res, vol. 22, pp.271-282, 2012.

[2] Matan hofree, John P Shen, et al “Network-based stratification of tumor mutations”, Nature method, vol. 10, no. 11, 2013

[3] The results are in whole or part based upon data

generated by the TCGA Research Network: http://cancergenome.nih.gov/
[4] Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561-D568 (2011).