TILD: A Strategy to Identify Cancer-related Genes Using Title Information in Literature Data

Jeongwoo Kim Hyunjin Kim Yunku Yeu Mincheol Shin Sanghyun Park*

Dept. of Computer Science, Yonsei University, 50 Yonsei-ro, Sinchon-dong, Seodamun-gu, Seoul 120-749, South Korea

{jwkim2013, chriskim, yyk, smanioso, sanghyun}@cs.yonsei.ac.kr
* corresponding author Tel: +82-2-2123-5714; fax: +82-2-365-2579

ABSTRACT

After genome project in 1990s, researches which are involved with gene have been progressed. These studies unearthed that gene is cause of disease, and relations between gene and disease are important. In this reason, we proposed a strategy called TILD that identifies cancer-related genes using title information in literature data. To implement our method, we selected cancer-specific literature data from the online database. We then extracted genes using text mining. In the next step, we classified into two kinds for extracted genes using title information. If genes are located in title, then they are classified as hub genes. In the contrast, if genes are located in body, then they are classified as sub genes which are connected with hub genes. We iterated the processes for each paper to construct the cancer-specific local gene network. In the last step, we constructed global cancer-specific gene network by integrating all local gene network, and calculated a score for each gene based on analysis of the global gene network. We assumed that genes in title have meaningful relations with cancer, and other genes in the body are related with the title genes. For validation, we compared with other methods for the top 20 genes inferred by each approach. Our approach found more cancer-related genes than comparable methods.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES –*biology and genetics.*

General Terms

Theory, Verification

Keywords

Text-mining; Cancer; Gene; Network; Analysis; Relation

METHODS

We proposed a method to identify cancer-genes using title information in literature data. We obtained abstract data from the PubMed. After preprocessing of the abstract data, we extracted genes in the literature. We then constructed local cancer gene network based on genes which are extracted by text-mining results. If the location of gene is title, then the gene is used as hub gene in local gene network. Otherwise, if the location of gene is text, then the gene is used as sub gene which is linked by hub gene. If the title of literature does not contain genes, then the literature is not used. The relations between hub genes and sub genes have weight. The weight is calculated based on location and frequency. The process of constructing local gene network is implemented for each paper. After constructing local gene network, we integrated all local gene networks to make global cancer gene network.

DTMBIO'14, November 7, 2014, Shanghai, China.
ACM 978-1-4503-1275-2/14/11.
http://dx.doi.org/10.1145/2665970.2665992

Scoring Function

We assumed that gene which is located in title has meaningful relations with cancer. We calculated a weight for each relation between genes using the frequency and location information. If title of literature has multi-genes then the relations between these genes have largest weight. On the contrary, the relations between sub genes which are included in text are removed in local gene network. The weight of relations between hub genes and sub genes is proportion to the frequency of sub gene. The frequency means the number of gene which is appeared in literature. The weight is calculated as follows:

$$\text{weight}(\text{h}, \text{h}) = 1$$

$$\text{weight}(\text{h}, \text{s}) = \frac{frequency(s)}{N}$$

$$\text{weight}(\text{s}, \text{s}) = 0$$

Here, *h* denotes the hub gene, and *s* denotes the sub gene. *N* indicates the number of all genes which are appeared in text area for each paper. The *frequency(s)* indicates the number of sub gene *s* in text area for each paper. After constructing local gene network, we generated global gene network based on nodes and genes of local gene network. The weight of overlapping relations is calculated by adding each weight of relations. After calculating the weight for each relation, we calculated a score for gene using degree centrality based on weight in the global cancer gene network.

Results & Conclusions

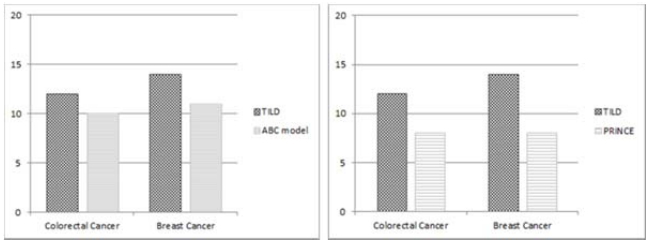

Fig. 1. Comparison TILD and comparable other methods the x-axis indicates diseases, and the y-axis indicates the number of cancer-related genes among the inferred top 20 genes. To validate candidate genes inferred by TILD methods, we performed literature validation. After verifying, we confirmed that candidate genes have also meaningful relation with cancer.

ACKNOWLEDGMENTS

This Research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea Government (MSIP) (2012R1A2A1A01010775).