SESE: Inferring disease-gene relationships using Second Sentence in biological literature

Jeongwoo Kim, Charndoh Bak, Inseop Kim, Seoyoung Kim, Junbum Cha, and Sanghyun Park

Abstract— **A vast number of researches which are involved in biology have been conducted, and large amount of literature data have been generated. These data are useful to discover biological knowledge. In biology, relationships between biological entities are important to discover cause of disease. To consider these issues, we proposed SESE method to infer disease-gene relationships using second sentence in literature data. To implement our method, we first obtain disease-specific literature data from PubMed. In the literature, we extract sentence which has a gene symbol and following sentence which has a pronoun. Analyzing these two type sentences, we calculate frequency based score for each gene. Using the score, we infer disease-gene relationships. For validation, we confirmed top k validation for genes inferred by SESE method and a comparable method. We demonstrated that our method is more suitable to infer disease-gene relationships than a comparable method. Additionally, we confirmed that second sentence is useful data to extract biological knowledge.**

I. METHODS

In our method, we use the first sentence to infer disease-gene relationships, and the second sentence is used to develop the relationships. We first gathered literature data which are involved in prostate cancer from the PubMed. To remove useless data in literature, we conduct a preprocessing. After preprocessing, we extract a sentence which includes gene symbol and followed sentence which includes pronoun. In case of followed sentence, we check meaning of the pronoun to extract interesting case that the pronoun indicates gene symbol. When the pronoun indicates gene symbol, the pronoun is counted as frequency of the gene symbol. The frequency indicates the number of appearance of gene symbol. In case of first sentence, we count the frequency for gene symbol without additional processing. After counting the frequency, we calculate a score for each gene to infer disease-gene relationships.

* This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2015R1A2A1A05001845).

J. Kim is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea;(jwkim2013@yonsei.ac.kr).

C. Bak is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea; (breakcd@nate.com)

I. Kim is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea; (insop90@naver.com)

S. Kim is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea; (youthskim@hanmail.net)

J. Cha is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea; (khanrc@yonsei.ac.kr)

S. Park(corresponding author) is with the Yonsei University, 50 Yonsei-ro, Shinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea (corresponding author to provide phone: +82-2-2123-7757; fax: +82 2 365 2579; e-mail: sanghyun@yonsei.ac.kr)

II. RESULTS AND DISCUSSIONS

In this section, we analyzed inferred Top 10 prostate cancer related genes. For several genes among the inferred top 10 genes by SESE method, we confirmed that they are involved in prostate cancer by answer set. For the other genes, we additionally validated relationships with prostate cancer by literature validation.

Table 1 shows inferred Top 10 gene by SESE method. In the table, the evidence indicates references of evidence to validate relationships between genes and prostate cancer.

Table 1. Top 10 genes inferred by SESE method

SESE	Evidence
AR	PGDB[4], KEGG[3], DDPC[1]
T	No evidence
EGFR	PGDB, DDPC
PTEN	PGDB, KEGG, DDPC, Sanger[5]
PCA3	PGDB, KEGG
ERG	Sanger
BRCA1	PGDB
STAT3	literature
TMPRSS2	PGDB, Sanger
BCR	No evidence

As shown in table 1, seven genes were validated by databases that they are involved in prostate cancer. On the other hand, two genes which include T and BCR were not validated by answer set and literature data.

The case of "STAT3", we found evidence that the gene has a relationship with prostate cancer based on literature published by Han Z et al. [2]. They presented relationships between STAT3 inhibition and efficacy for prostate cancer treatment in their experiments.

REFERENCES

[1] Maqungo M., Kaur M., Kwofie SK., Radovanovic A., Schaefer U., Schmeier S., Oppon E., Christoffels A., and Bajic VB. DDPC: dragon database of genes associated with prostate cancer. Nucl Acids Res 2010. http://dx.doi.org/10.1093/nar/gkq849.

[2] Han Z., Wang X., Ma L., Chen L., Xiao M., Huang L., Cao Y., Bai J., Ma D., Zhou J., and Hong Z. Inhibition of STAT3 signaling targets both tumor-initiating and differentiated cell populations in prostate cancer. Oncotarget. 2014. Sep 30;5(18):8416-28.

[3] KEGG: Kyoto Encyclopedia of Genes and Genomes. DOI=http://www.genome.jp/kegg/

[4] Li LC., Zhao H., Shiina H., Kane CJ., and Dahiya R. PGDB: a curated and integrated database of genes related to the prostate. Nucl Acids Res 2003;31(1): 291-3.

[5] Wellcome Trust Sanger Institute <http://www.sanger.ac.uk>