Extracting Genotype-Phenotype Relationships From Literature Using Natural Language Processing

Nate Sutton[1], Graciela Gonzalez, PhD[1]

[1]Department of Biomedical Informatics, Arizona State University, AZ

INTRODUCTION

An abundance of publications on genotype-phenotype findings continue to be generated. Finding and reviewing those publications can be quite a time consuming task for researchers.

Natural language processing (NLP) can help those researchers to collect and find meaning within those publications. One way that NLP can help inform researchers of meaning in those publications is by aggregating and summarizing the findings extracted from that literature.

METHODS

A variety of methods are included in the NLP analyses. Initially, literature in the form of XML is automatically retrieved from Pubmed and Pubmed Central in the form of XML. Next, sentence detection and tokenization is performed. Open source software assists with those tasks.

Named entity recognition (NER) is performed with a combination of dictionary-based, regular expression, and machine learning methods. The machine learning approach used was conditional random fields. Some entities have been tested with more than one of those NER methods to find performance differences in applying the alternate methods.

Once the entities have been identified within text they then are tested to find if relationships can be identified amongst them. Several different methods are used in genotype-phenotype relationship identification. Some of those methods are identification of co-occurring entities, linguistic rule based methods, and machine learning.

Different types of scoring criteria have been considered. Currently only exact matches of all entities within gold standard relationships is counted as correctly identified relationships. However, while that criteria can be reported, findings with some inexactness but are also informative about relationships may be useful to report.

RESULTS

Preliminary evaluation F1-scores based on particularly limited training and testing data for the detection of genetic markers range from 0.36-0.93. Also, relationship detection showed an accuracy score of 0.2. However, in general, those scores simply represent a proof of concept that the system works. More training and testing data can substantially improve scores. That data can be further generated through annotations.

Association between SMAD7 and colorectal cancer

We evaluated selected SNPs in three replication sample sets (7,473 cases, 5,984 controls) and identified three SNPs in SMAD7 (involved in TGF-b and Wnt signaling) associated with CRC (1).

Association between rs6983267 and colorectal cancer

The most strongly associated SNP ($P = 1.72 \times 10^{-7}$, allelic test) was rs6983267 at 8q24.21 (2).

Figure 1. Examples of findings that are of interest to extract with key terms highlighted.

CONCLUSIONS

NLP is capable of being a valuable method for efficiently retrieving genotype-phenotype findings. That retrieval can provide timely and useful information about those findings to researchers. The performance of NLP at identifying those findings can improve through further experimentation and integration of multiple methods that have been described in publications.

REFERENCES

1. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet. 2007 Nov;39(11):1315-1317.
2. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. 2007;39(8):984-988.

ASU biomedicine ARIZONA STATE UNIVERSITY
DEPARTMENT OF BIOMEDICAL INFORMATICS

Printing Funded In Part By GPSA