Using Natural Language Processing to Automatically Extract Alzheimer's Disease Related Genotype-Phenotype and Pharmacogenomic Findings

Nate Sutton[1], Graciela Gonzalez[1]

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

Introduction

Multiple gene variants in addition to variants in APOE have been associated with many cases of Alzheimer's disease (AD). Collecting and identifying such findings that have been published about AD can be a time consuming task for researchers.

Natural language processing (NLP) can help those researchers to collect and find meaning within those publications. An initial version of a NLP system that can extract genotype-phenotype findings specifically about AD and also other phenotypes has be created. Further, development has been performed on a system that can automatically connect those extracted findings to direct or indirect genetic marker-drug (pharmacogenomic) relationships.

Materials and Methods

- Currently, the same machine learning technique that the BANNER system uses, conditional random fields (CRFs), are used for detection of all entities including genes and relationships (1). Examples of relationships that the system is designed to extract are shown in figure 1.
- Integration is planned of national library of medicine's MetaMap for diseases, medical subject headings for study types and species, and BANNER to improve named entity recognition of genes (2,3).
- Pharmacogenomic connections with extracted findings are automatically made by a program that creates specialized queries to a downloaded pharmacogenomic database.
- A planned feature is to allow users to retrieve either direct or indirect pharmacogenomic relationships with extracted findings. Indirect relationships can have the advantage of being novel putative relationships that are not yet reported in literature. A graphical representation of the relationship types is shown in figure 2.
- Work has been done on a program that maps multiple types of extracted markers to drugs in the pharmacogenomics database. The series of operations that program performs are shown in figure 3.

Association between CLU, CR1 and Alzheimer's disease

Genome-wide association study identifies variants at *CLU* and *CR1* associated with Alzheimer's disease. (4)

Association between MS4A and Alzheimer's disease

The membrane-spanning 4-domains, subfamily A (MS4A) gene cluster contains a common variant associated with Alzheimer's Disease. (5)

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

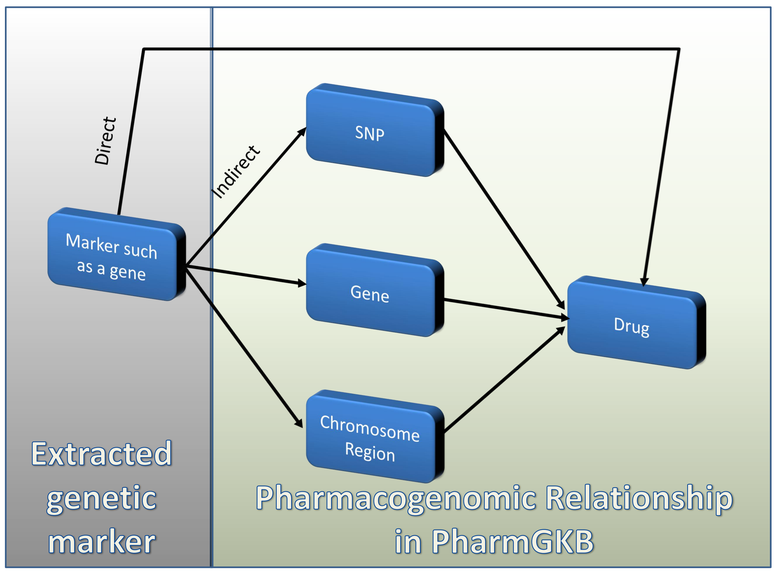

Figure 2. The process of making direct and indirect connections between extracted markers and drugs in PharmGKB.

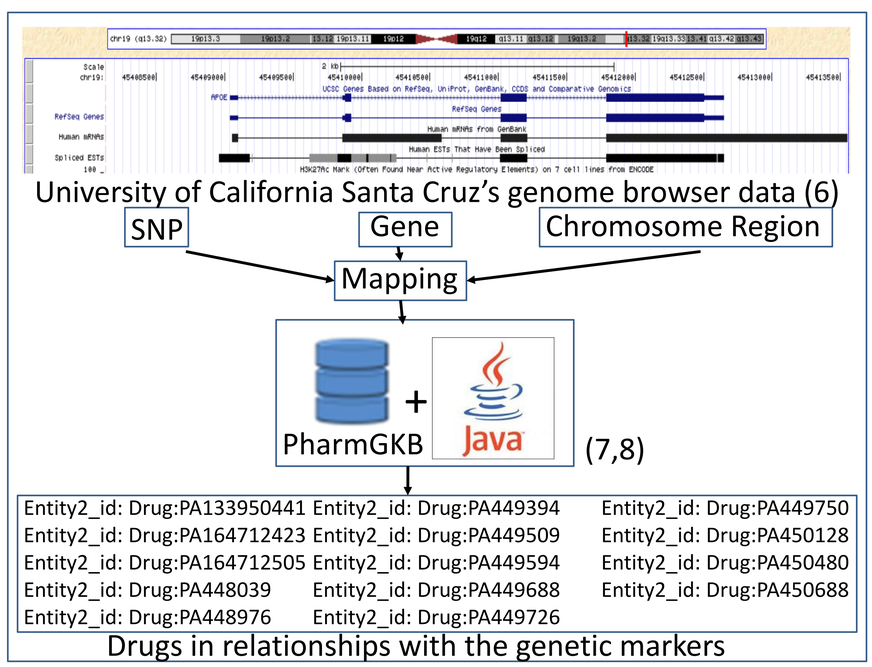

Figure 3. The sequence of events involved in mapping extracted markers to genes. Queries are then made of those genes with a subset of PharmGKB data to find pharmacogenomic relationships.

Results

- Preliminary F1 evaluation scores: SNPs: 0.93, Genes: 0.36, chromosome regions: 0.44, evidence terms: 0.42, and relationships: 0.20.
- However, in general, those scores simply represent a proof of concept that the system works. At present, only particularly limited training and testing data has been used to generate the evaluation results.
- The automated drug relationship findings showed accurate results when they were evaluated manually.
- 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.

Conclusions

- Valuable approaches to efficiently retrieve DNA-based study results about Alzheimer's disease are supplied with in the NLP system that is presented here.
- Some additional future work on the system includes statistical summarization and visualization components that help inform users of the results that the system extracts.
- Connecting pharmacogenomic relationships to the genotype-phenotype findings also provides particularly useful information to researchers that can be used to help guide future research that they undertake.

References

1. Robert Leaman, Graciela Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition [Internet]. 2008 [cited 2011 Jun 8];Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.137.8686
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;:17-21.
3. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc. 2000 Jul;88(3):265-266.
4. Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 2009 Oct;41(10):1094-1099.
5. Antunez C, Boada M, Gonzalez-Perez A, Gayan J, Ramirez-Lorca R, Marin J, et al. The membrane-spanning 4-domains, subfamily A (MS4A) gene cluster contains a common variant associated with Alzheimer's Disease. Genome Med. 2011 May 31;3(5):33.
6. Human chr19:45,409,039-45,412,649 - UCSC Genome Browser v251 [Internet]. [cited 2011 Jun 7];Available from: http://genome.ucsc.edu/cgi-bin/hgTracks
7. sun-java.jpg (JPEG Image, 300x300 pixels) [Internet]. [cited 2011 Jun 7];Available from: http://www.java-entrepreneur.com/wp-content/uploads/sun-java.jpg
8. The Pharmacogenomics Knowledge Base [PharmGKB] [Internet]. [cited 2011 Jun 7];Available from: http://www.pharmgkb.org/

ASU biomedicine
ARIZONA STATE UNIVERSITY
DEPARTMENT OF BIOMEDICAL INFORMATICS