# Text mining in cancer gene prioritization

Text mining has been an efficient tool for predicting the genetics of diseases. During the process, it will help the experts to identify the qualified gene candidates for the further analysis. Furthermore, the prioritization of cancer genes means that sorting the candidate genes based on their similarities, which helps a lot on reducing the research cost. The researchers had developed lots of gene prioritization tools and data sources, they are mainly about gene sequences, differential expressions, protein domains, and pathways. So, we will discuss several text mining methods for gene prioritization.

Firstly, we need to learn cancer genes' data sources. Since the phenotypes of cancers are relevant with full levels of the Omic hierarchy including genomics, transcriptomics, epigenomics, proteomics, and metabolomics. Also, these data used many ways to express themselves, such as numerical expression, narrative text and sequence read. These data sources have been separated into several types because of different information exchange and Omic hierarchy, and they had been classified based on their primary utility. Most of the data sources belong to text and literature, which means that text mining could be largely implemented in the gene prioritization.

The first method is prioritization with keyword search text mining. This method is more sensitive to keyword matching when retrieving literature or text for responding query terms. In purpose of pre-classifying and shortening the running time, the indexing will preprocess this original text with data sources. We learned that gene seeker depends on gene positional and expression data. Here, prioritizer focuses on specified protein interactions and pathway data, then it constructs the gene network according to Bayesian integration. Then users will be requested to provide keywords relating to their desired features, which will be used to search the literature and find protein domain terms. The tool will combine other data sources to generate the relative scores and all source specific scores will be weighted to generate a final grade.

The second method is prioritization with statistical text mining. To improve prediction accuracy, this method utilizes a predetermined distribution or a Bayesian model to fit the data. GRAIL is a text mining approach that not only discovers illness genes, but also their biological connections, in the same manner that disease genes are connected to disease pathways. GRAIL connected seed chromosomal locations to GWAS-identified SNP involved seeds, lowering the likelihood of knowledge contamination. Genie is a bibliology-based prioritizing tool that searches MEDLINE for relevant gene abstracts and assesses homology. MetaRanker took into account GWAS SNP-phenotypic connections, PPI networks, linkage analysis data, gene expression, and user data as well.

The third method is prioritization with text mining using ontology structure. Gene prioritization techniques in this toll dive deeper into the ontology structure to assess semantic similarity between concepts with more precision. To link literature text mining with gene expression data, the EVOC anatomical ontology was employed. The bridging disease gene pairs were then prioritized using heuristic sequencing scores. MedSim employs the GO term to annotate genes' functions, and it runs many configuration tests to see if homologous or interacting genes are included. The SimRel score was used to determine semantic similarity, which took into consideration the differences, commonness, and particularity of GO words.

Text mining is a major technique because of the large volume and frequency of biological literature and narrative text across several distinct data sources. We looked at and classed gene prioritization text mining techniques based on the more sophisticated models deployed, noting that they have evolved more slowly than other aspects of gene prioritization that analyze structured data. Text-network translation to offer finer semantic information and pathway prioritization to provide extra mechanism knowledge are two next possibilities we've identified. It would wildly be implemented for cancer genes prioritizations soon.

References,

Luo, Y., Riedlinger, G., & Szolovits, P. (2014, January 1). Text Mining In Cancer Gene. And Pathway Prioritization. PubMed Central (PMC). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216063/.

Yu, S., Tranchevent, L., & Moor, B. (2010, January 14). Gene Prioritization And. Clustering By Multi-view Text Mining - BMC Bioinformatics. BMC

Bioinformatics. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-28.