Biomedical literature mining contextualizes genes and identifies novel therapeutic gene targets in cancer research.

Peng-Chan Lin MD., Ph.D.

Abstract

Developing a biomedical-explainable and validatable text mining pipeline can help in cancer gene panel discovery. We create a pipeline that can contextualize genes by using text-mined co-occurrence features. We apply Biomedical Natural Language Processing (BioNLP) techniques for literature mining in the cancer gene panel. A literature-derived 4,679 × 4,630 gene term-feature matrix was built. The best accuracy for predicting two different gene panels in different machine learning models, including MSK-IMPACT (Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets), and the Oncomine cancer gene panel, is 0.959 and 0.989, respectively. The use of text-mined co-occurrence features can contextualize each gene. We believe that we can predict the genes for cancer discovery. In conclusion, this study highlights the importance of biomedical literature mining in gene panel discovery and interpretation. The platform could provide an opportunity to construct a gene recommendation and annotation system for precision medicine.