A Novel Computational Approach for Reducing False Positives in Text Data Mining

Authors: Noah Yasarturk, Surajit Bhattacharya

Faculty Sponsor: Dr. Daniel N. Cox Neuroscience

Elucidating molecular genetic mechanisms regulating complex biological processes has benefit from the advent of large-scale bioinformatics efforts that provide tools for mining genetic interactions and functions in a tissue or cell type of interest. Among these bioinformatics tools is the use of text data mining algorithms, which are computational strategies for extracting biological data on gene interactions and gene ontologies for a queried gene from the published scientific literature. One major problem with these algorithms is a high false discovery rate (FDR) whereby genetic interactions or gene functions are sometimes incorrectly reported for a given gene because most of these algorithms look at whether there is a presence of the queried genes and the functions or the interaction, but do not check the connections between the same. We hypothesize that FDR from text mining for a given query gene can be reduced by implementing algorithms that more rigorously assess data extracted from text mining.

Here we construct an algorithm wherein we first manually curate literature and look at the most frequently used terms that connect a query gene to a function or an interaction, and then use this method to construct a computational algorithm which would 'read' in an article and search for the distance between the queried words and the functions/interactions, using nearest neighbor algorithms. We are presently building a desktop based tool using this algorithm, and will initially focus on the rich primary literature and data resources of the model organism, *Drosophila melanogaster* via the FlyBase database.

Keywords: text data mining; bioinformatics; computational biology; false discovery rate; databases; *Drosophila*; genetic interactions; gene functions

Embargo: We will be embargoing this work for this presentation this semester.