CALIFORNIA SCIENCE & ENGINEERING FAIR
2019 PROJECT SUMMARY

Name(s)

Ayush Alag

Project Number

S0803

Project Title

Computational DNA Methylation Analysis of Food Allergy Yields Novel 13-gene Signature to Diagnose Clinical Reactivity

Abstract

Objectives

Current blood and skin tests are inaccurate (50-60% false positive rate) in distinguishing true food allergies (FA) from oral sensitivities. As a result, life-threatening Oral Food Challenges (OFC) are used, which has resulted in patient mortality and over-diagnosis of FA. I sought to create a highly-accurate diagnostic classifier of FA from blood sample (safer than an OFC) epigenomic data. I also sought to use a purely data-driven methodology, which would be rendered applicable to other diseases. Lastly, I sought to find biological associations with FA.

Methods

Working by myself on a dataset publicly available on Gene Expression Omnibus, I coded in Java (with Weka ML library) to develop a computational framework for feature selection and classification. My methodology was based on Sequential Forward Selection and ensemble classification methods. I later used the Illumina BaseSpace Correlation Engine to find gene and pathway associations for the diseases I found. I also used Gene Ontology Enrichment Analysis, Princeton University s Generic Gene Ontology Term Mapper, REVIGO, and NAVIGO, which are all publicly available, to find representational biological terms associated with the genes I found.

Results

An unbiased feature-selection pipeline was created that narrowed down 405,000+ potential CpG biomarkers to 18. Machine-learning models that utilized subsets of this 18-feature aggregate achieved perfect classification accuracy on completely hidden test cohorts. Ensemble classification was also shown to be effective for this High Dimension Low Sample Size (HDLSS) DNA methylation dataset. The 18-CpG signature mapped to 13 genes, on which biological insights were collected. Notably, many of the FA-discriminating genes found in this study were strongly associated with the immune system, and seven of the 13 genes were previously associated with FA.

Conclusions

I implemented an efficient feature-selection algorithm that found a condensed list of strong CpG biomarkers. I replicated the perfect classification found in previous works but with a much smaller CpG set (by Occam Learning, simpler models are preferable), and also with unbiased k-fold cross-validation accuracy measurement. Furthermore, the methodology I used was completely data-driven and generalizable to other diseases. I also found novel genes associated with FA. I am the sole author of this paper s publication in PLOS One, and it is currently in the minor edits stage.

Summary Statement

I created purely data-driven and highly-accurate machine learning models to perfectly classify true food allergies (as opposed to milder sensitivities), and in this process I found genes and biological pathways associated with the disease.

Help Received

Working independently at home, I consulted Dr. Joseph Hernandez from Stanford University for feedback on my paper, and he also advised me on collecting biological insights.