CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Elleen Kim

Project Number
38492

Project Title
Using Machine Learning Algorithms to Investigate Genetic Burden of Scoliosis from Candidate Genes & GWAS Studies

Abstract

Objectives/Goals
The engineering goal was to create programs that cluster characteristics of Scoliosis candidate genes and prioritize pairs of Scoliosis patient DNA sequences to investigate the causes of Scoliosis.

Methods/Materials
A database of candidate genes was compiled via literature review of published papers. Candidate genes were grouped by running their numerical characteristics through KNN and K-Means clustering algorithms. Scoliosis patient DNA sequences from public databases were run through customized local and global alignment programs and through BLAST.

Results
Results from the KNN program provided an upper bound and lower bound threshold to quantify the validation test scores of clustering, allowing patient data to be quantified by quality according to this threshold. KNN and K-Means are meaningful analyses of gene groups that can elucidate underlying similarities of Scoliosis candidate genes while producing consistent results within the margin of error. Results from alignment provided a prioritized list of Scoliosis DNA sequences.

Conclusions/Discussion
In conclusion, this project is a proof of concept experiment that shows that machine learning algorithms can be applied to genetic data. Prioritized DNA sequences can be used to further study the association to Scoliosis. With the falling cost and rising prevalence of GWAS studies, this prioritization can be used to quantify the influx of GWAS data for various diseases. Alignment programs also provided meaningful differences in best alignment scores between patient DNA sequences to use as a starting point for further research on what possible genetic variations or phenotypic characteristics determine best alignment. The machine learning clustering algorithms produced a range in validation test measures to quantify genetic datasets. Future identified candidate can be assessed using this metric. In line with our experimental goals, identified clusters or groups of candidate genes also serve as a foothold for further investigation on potential relationships or patterns between the genes and Scoliosis.

Summary Statement
This project customizes machine learning clustering algorithms and DNA alignment algorithms to study Scoliosis candidate genes and patient DNA, and creates a metric to quantify the quality of GWAS datasets in other disease applications.

Help Received
I designed and built the programs by myself. I got help in understanding the algorithms from Dr. Neil Sarkar (Department of Biomedical Informatics, Brown University) and Sophie Kim (Department of Humanities and Sciences, Stanford University).