CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)
Shivam Singhal

Project Number
38393

Project Title
iDetect: A Machine Learning Algorithm for Non-Invasive Cancer Diagnosis through Epigenetic Biomarker Identification

Abstract

Objectives/Goals
Cancer remains a leading cause of death in today's world, only treatable if caught early. Epigenetic alterations are newly-discovered biomarkers that can facilitate early diagnosis. They are particularly attractive as a diagnostic tool due to their stability, frequency, and non-invasive accessibility in bodily fluids, such as blood plasma, in the form of cell-free DNA (cfDNA). The objectives of this project were to locate epigenetic alterations in cfDNA sequences obtained from the blood plasma of cancer patients and to map these sites to CpG islands (CpGI) to identify biomarkers for accurate and non-invasive cancer diagnosis using a machine learning algorithm.

Methods/Materials
This project was conducted in four steps: 1. Determining methylation in cell-free DNA sequences relative to the reference genome using the Bismark tool in Perl script. 2. Mapping the methylated and unmethylated sites to the CpG islands, which are important genomic regions for cancer detection. 3. Selecting features that indicate cancer presence and training the machine learning algorithm with tissue data obtained from the TCGA database to look for these specific features when making predictions for different types of cancer. 4. Testing the accuracy of the algorithm through cfDNA samples obtained from another researcher.

Results
The alignment of the cell-free DNA sequences in Bismark showed that different methylation levels are present, which enabled successful CpGI mapping. A positive correlation between differential methylation levels of cell-free DNA and tissue DNA samples allowed for the use of tissue data in the machine learning process, as well as validating the effectiveness of alterations in cell-free DNA as biomarkers for cancer. The algorithm was able to differentiate between cancerous DNA and the non-cancerous (control) DNA with an 85 percent sensitivity and 67 percent specificity.

Conclusions/Discussion
The program was able to accurately differentiate between different types of cancers for which cfDNA data was available. This algorithm can also be applicable to a host of other conditions in which identifiable differences in methylation have been reported, including neuropsychiatric disorders and cardiovascular diseases. In addition, this non-invasive technique can be extended to more advanced prenatal tests for unborn fetuses.

Summary Statement
This project developed a machine learning algorithm which considers methylation patterns in cell-free DNA to predict cancer presence accurately, efficiently, and non-invasively.

Help Received
I would like to thank Dr. Jiang for providing the cell-free DNA data. I would also like to acknowledge my computer science teacher, Mr. Steinke, for providing me with constant encouragement and my parents for their support throughout the project.