CALIFORNIA STATE SCIENCE FAIR
2013 PROJECT SUMMARY

Name(s)

Andrew C. Jin

Project Number

S0508

Project Title

Breast Cancer Prognosis through Gene Expression Profiling and Tumor Morphology

Abstract

Objectives/Goals

In a world where 20% of breast cancer survivors suffer recurrence, accurate tumor prognosis and recurrence prediction are becoming increasingly crucial in ensuring that patients receive proper attention and treatment; for example, a tumor correctly identified as aggressive can be given a more potent form of treatment to ensure that recurrence does not occur.

The purpose of this project was twofold. First, to develop a Support Vector Machine (SVM) prognostic model that accurately predicts whether a breast cancer patient will experience a recurrence within three years after treatment. Second, to identify highly predictive features in the model, which allows for the discovery of characteristics that play integral roles in tumor proliferation and recurrence; this aids in the discovery of potential biomarkers and therapeutic targets.

Methods/Materials

Three types of features were explored when creating the SVM model: gene expression values of individual genes, gene expression analyzed holistically over gene sets, and morphologic and structural features from tumor slide images. A gene set is a group of functionally related genes accounting for a particular cellular process (e.g. apoptosis genes, inflammation genes, or cell cycle genes).

Results

After going through supervised learning on the training set, the integrative SVM model incorporating all three feature types (gene expression, gene set, and image features) yielded an accuracy rate of 86.7% when predicting recurrence outcome for the validation set. The gene set model (80% accuracy) and image model (73.3% accuracy) also displayed substantial predictive ability, but the gene expression model performed poorly (66.7% accuracy).

Conclusions/Discussion

Within the optimal subset of 132 features used in the model, features such as the Symporter Activity Gene Set and Actin Filament Binding Gene Set exhibit significant causal relations to breast cancer recurrence. These newly implicated features can be biologically explored to better understand the mechanisms of cancer.

Ultimately, although computer models may never replace pathologists, they prove to be extremely effective tools in tumor grading and prognosis.

Summary Statement

I developed a prognostic model based on tumor morphology from slide images, mRNA expression values, and gene set alterations across entire cellular functions; the novel, integrative framework accurately predicts breast cancer recurrence.

Help Received

Mentored by Dr. Andrew Beck of the Harvard Medical School; advice from research associate Nicholas Knoblauch; advice from teacher sponsor Mr. Chris Spenner