CALIFORNIA STATE SCIENCE FAIR
2015 PROJECT SUMMARY

Name(s)
Simon L. Kuang

Project Number
S1414

Project Title
Predicting Chromosome Conformation from Epigenetic Features

Abstract

Objectives/Goals
The objective is to predict the results of 40kb HinDIII Hi-C experimentation on the mouse ESC genome from 21 documented mouse epigenetic features.

Methods/Materials
Around 200 million 22-dimensional datapoints were drawn from the NIH histone database and from the Hi-C paper. Characteristics of the data required development of a novel machine learning meta-algorithm using inverse probability Metropolis-Hastings sampling, bagging with 100 independent samples, single-layer perceptrons optimized using Levenberg-Marquardt training, and intelligent weighting based on in-sample prediction correlation. Chromosome 1 was used for training, chromosome 2 for validation, and chromosomes 3 to X for testing.

Results
The model's predictions scored a Pearson correlation coefficient between 0.7 and 0.8 on all chromosomes. In-sample RMSE and out-of-sample RMSE decreased on all predictors over the 5 training epochs, and in-sample correlation and out-of-sample correlation increased on all predictors over the 5 training epochs.

Conclusions/Discussion
My method performs extremely well on this unusual dataset. In the next decade, spatial distance will be more insightful, cheaper and faster to compute, and easier to analyze in clinical situations than the DNA sequence itself.

Summary Statement
A novel machine learning meta-algorithm is necessary and effective for studying the 3-dimensional conformation of the mammalian genome.

Help Received
Used computing facilities at PSU Center for Comparative Genomics and Bioinformatics