Mining Interactions in Candidate Genes for Hypertension

Ruteja Barve

Mentor: C. Charles Gu, Ph.D.

Abstract:

Hypertension (HT) is an important public health problem affecting more than 60 million people in the US and over a billion people worldwide. Essential hypertension (high blood pressure with no obvious causes) accounts for 95% of the cases. The biological process of blood pressure regulation involves multiple physiological pathways, each of which may be affected by multiple genes. Therefore, it is expected that multiple factors (genetic and environmental) may collectively predisposing individuals to susceptibility or resistance to HT. The underlying interaction of the involving factors is extremely complex. Many studies reported association of HT to a slew of candidate genes but the replication of positive findings is scarce. One major reason is believed to be the lack of proper modeling of interaction effects. The aim of this project is to evaluate by analyzing real data, the effect of different ways of modeling interactions on the chance of identifying important factors for HT.

An existing dataset from the GENOA study (Genetic Epidemiology Network of Arteriopathy, part of the Family Blood Pressure Program) was used for this project. The dataset contains a sample of 3075 hypertensive patients and 1074 normotensive controls (in African, Caucasian and Hispanic Americans), all genotyped for 127 SNPs in 21 candidate genes on Chromosome 2 where linkage to HT was previously detected. We compared three strategies: (1) single-SNP analysis (no interaction modeling), (2) brute force classification of HT status by naïve Bayes and CART to all 127 SNPs, and (3) performing variable selection by Random Forest analysis to form best predictive model of HT. The application of classification trees in (2 & 3) implicitly accounts for interaction effects. We considered different models by including, in addition to the SNPs, risk factors of HT (BMI, triglyceride, LDL, HDL, insulin and glucose) and use of hypertensive medications to accommodate potential GxE interactions. Using estimated prediction error as performance measure, we found that Strategy 3 consistently selected top-ranked SNPs and covariates and performed much better than the other two with a prediction error of 11-17%. We found that when we considered only the SNPs, 2 genes (IL1B and SLC4A5) were ranked as important in all 3 race groups. As more covariates and environmental factors were added to the model, different patterns of combinations of the variables emerged, which likely reflect the underlying interaction effects driving the HT outcome observed in the different race groups. While logistic regression with interaction terms found no significant interaction effects, we also performed validation analysis using MDR to directly test for interactions and found various significant synergistic SNP-SNP interactions. In summary, our analyses showed that performing variable selection using a statistical learning method such as the Random Forest seems to be a viable strategy to select important risk factors for HT while taking into account of important underlying interaction effects. Mining the interactions among the 21 candidate genes this way helped us to find the promising roles of IL1B and SLC4A5 and their interactions in HT.