



# CALIFORNIA STATE SCIENCE FAIR 2012 PROJECT SUMMARY

<b>Name(s)</b> <b>James J. Thomas</b>	<b>Project Number</b> <b>S1214</b>
<b>Project Title</b> <b>A Predictive Bayesian Network Model of Alcohol Dependence Based on Genetic and Demographic Factors</b>	
<p style="text-align: center;"><b>Abstract</b></p> <p><b>Objectives/Goals</b> We sought to develop the first predictive model of alcohol dependence and analyze the model to pinpoint the biological mechanisms that underlie the onset of the disease.</p> <p><b>Methods/Materials</b> We began with a dataset containing demographic and genetic information for 3,776 individuals, about half of whom were alcoholics. We invented a generalizable approach based on linkage disequilibrium to eliminate redundant and uninformative genetic factors from the dataset. A Bayesian network-based predictive model was learned from the reduced dataset using the Java-based WEKA machine learning library. We also invented a method to analyze any Bayesian network model of a disease to identify genes and high-level biological concepts (e.g. pathways) that are likely to be associated with the disease. We implemented this method as a Java software package and, on a supercomputer, applied it to our Bayesian network for alcoholism.</p> <p><b>Results</b> The area under the receiver operating characteristic curve (AUROC), a measure of predictive power, is a very high 92% for our model. Many of the genes and biological concepts that were deemed likely to play a role in the onset of alcoholism have already been tied in some way to the disease in the literature, validating both our novel method for model analysis and the model itself. Other findings do not seem to have been documented previously in the literature and are excellent candidates for further research.</p> <p><b>Conclusions/Discussion</b> We have developed a highly accurate model to determine the probability that a patient will develop alcohol dependence given his or her genetic and demographic information. Examination of the model's network structure indicates that a major reason for the model's success is a synergy between demographic and genetic factors -- the presence of certain demographic factors increases the likelihood that certain genetic factors will cause alcoholism. We have determined several new links between genes and biological concepts and the onset of alcoholism, many of which may aid in the development of preventative treatments for the disease (drugs that target certain pathways, gene therapy measures, etc.). Finally, all of our methods, including our novel algorithms for genetic feature selection and analysis of Bayesian network models of diseases, have been validated by the success of our model and can be used in the future to construct and analyze models of other diseases.</p>	
<b>Summary Statement</b> Using Bayesian statistics, we developed the first model that can predict the onset of alcoholism in individuals and analyzed it to better characterize the biological mechanisms of the disease.	
<b>Help Received</b> Project was conducted under the guidance of Dr. Amin Zollanvari, Research Fellow, Children's Hospital Informatics Program	