MODELING EPISTATIC INTERACTIONS USING MACHINE LEARNING TECHNIQUES

Sky Adams (Brown University) & Regie Felix (CSU San Bernardino)
Mentors: Petko Bogdanov and Nick Beck PI: Professor Ambuj Singh (Department of Computer Science)

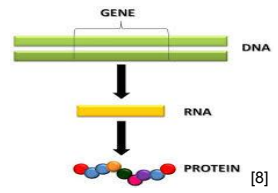
Supported by : **NSF - REU** IIS- 0917149 and IIS-1219254

Motivation: The regions of the genome involved in many complex genetic diseases remain unknown. There are significant interactions within the genome that greatly influence the phenotype of an organism, specifically between single nucleotide polymorphisms (SNPs) and gene expression levels. The detection of these interactions is difficult due to the huge set of SNPs and genes and the small marginal effect of each marker. Our goal is to identify the meaningful SNP-Phenotype interactions and SNP-Gene-Phenotype interactions for both synthetic and real data sets.

Biology Background:

SNP location

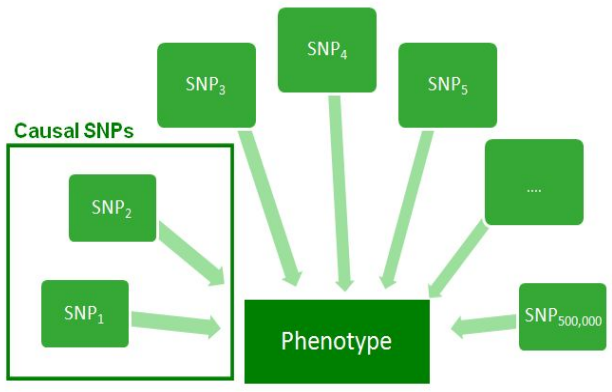
DNA sequences: organism 1: ...ATTCATTG... genotype at this SNP: A
organism 2: ...ATTCCTTG... genotype at this SNP: C



[8]

DNA is transcribed to RNA. The amount of RNA present from a particular gene is the **expression level** of that gene.

Two Challenges:



1. Associated genomic factors may not have a large marginal effect, making it necessary to evaluate groups of factors together instead of each factor individually.

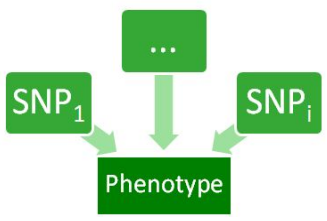
2. The search space of all possible sets of genomic factors is exponential in the number of factors, making an exhaustive search infeasible.

Machine Learning Methods:

Bayesian Networks (BN)	Decision Trees	Multifactor Dimensionality Reduction (MDR)	Consistency subset (CS) analysis with resampling	Support Vector Machine (SVM)
• BDeu, AIC, BIC-- score networks based on how well they fit the data but penalize larger networks [2] • Markov chain Monte Carlo (MCMC) search calculates the posterior probability that each genomic factor is associated with the phenotype[6]	• Each node is a genomic factor and each branch is a possible value for the factor. Leaves show the predicted phenotype of the samples with the corresponding genotype [3] • Classifies instances by threading them down the tree from the root to a leaf node	• Based on a case and control ratio, a combination of genomic factors is either high-risk if they exceed threshold T or low-risk if they are below the threshold [2]	• Looks for sets of the genomic factors which can correctly classify samples • To prevent finding local maximum and to avoid overfitting, we resampled the samples k times, performed CS analysis, and then chose the smallest set from the k+1 resulting sets of genomic factors [4]	•Classifies the data by defining a hyperplane between the two different classes [5] •Ranks genomic factors by their ability to classify samples [7]

Part I: SNP - Phenotype Interactions

Our Model

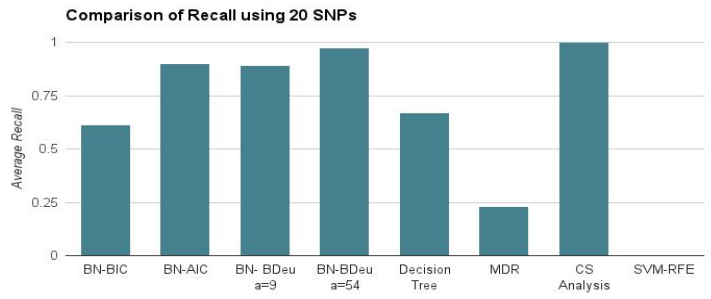


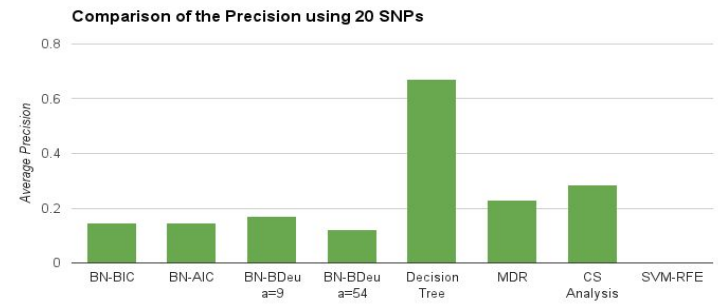
Our Process

Evaluation results considering only SNPs and not gene expression levels

Data type:	Synthetic
Number of SNPs:	20 (two are causal)
Number of samples:	200
Phenotype:	Binary disease state

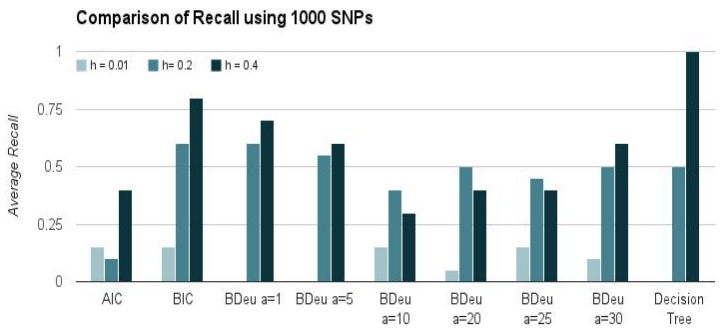
Note: SVM-RFE does not return a set of SNPs, but rather sorts the SNPs in an order that reflects their relatedness to the phenotype. The rankings showed no preference for the two causal SNPs.



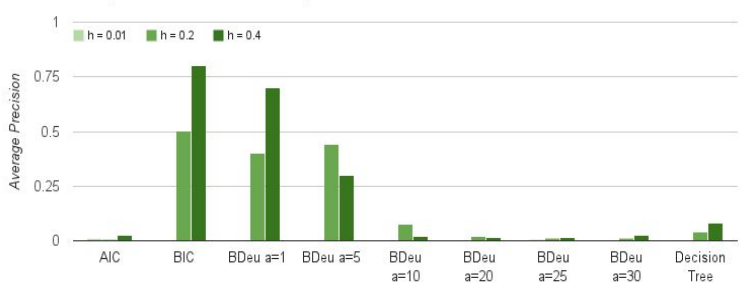


Data type:	Synthetic
Number of SNPs:	1000 (two are causal)
Number of samples:	200
Phenotype:	Binary disease state

Note: CS Analysis and MDR are not included. CS Analysis with resampling is computationally infeasible when there are 1000 SNPs. MDR results in zero precision and recall because it does not find the correct causal SNPs.

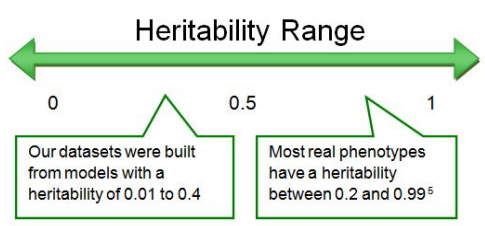


"h" = Heritability



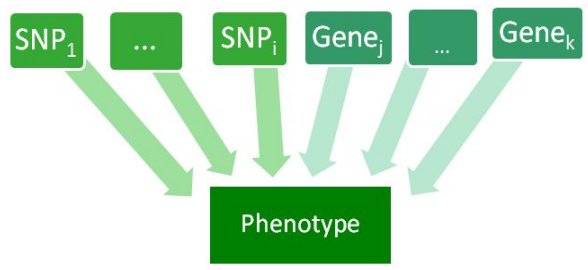
Analysis of Evaluation Results

If the method is more sensitive to SNPs with small effects, it will result in:	If the method avoids overfitting the model, it will result in:
High recall Low precision	High precision Low recall



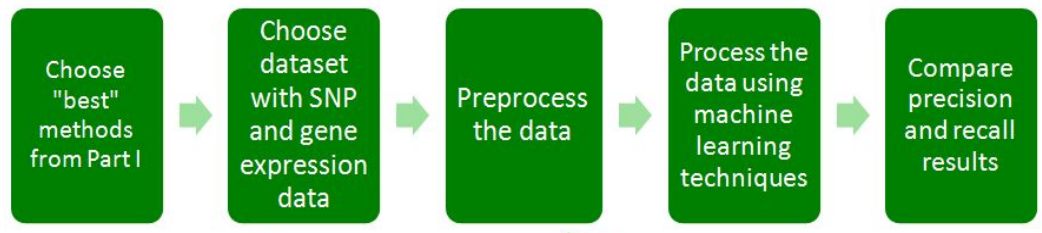
Part II: SNP - Gene - Phenotype Interactions

Our Model

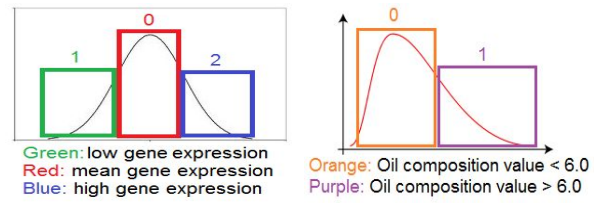


A model which incorporates gene expression levels is more accurate than a SNP-only model because expression levels account for a significant portion of phenotypic variance.

Our Process

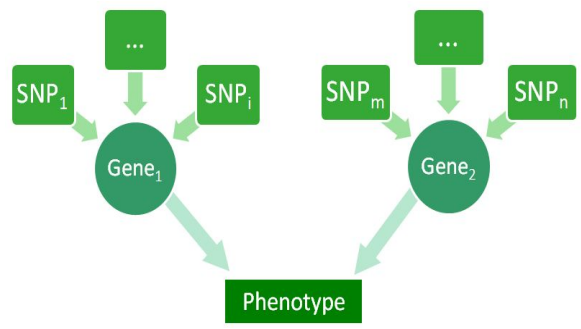


Instead of using the raw data, which has continuous values for gene expression and oil composition, we determine thresholds that categorize each sample by whether it is above or below each threshold.



Future work

- Analyze the SNP-Gene-Phenotype model on the Maize data:
  Bayesian Network (using BDeu, AIC, BIC), Decision Trees, etc..



- Develop a method that produces a model in which groups of SNPs affect the expression levels of genes and those genes affect the organism's phenotype or disease state.

- Apply our method from Part II to other realistic datasets

- Improve the methods to achieve more reliable results (higher precision and recall values)

References/Citations

[1] Genome-wide association study dissects the genetic architecture of biosynthesis in maize kernels by Hui Li, et. al. (2013)
[2] Learning genetic epistasis using Bayesian network scoring criteria by Xia Jiang, et. al. (2011)
[3] Machine Learning by Tim Mitchell (1997)
[4] Model search and inference by bootstrap "bumping" by Tibshirani and Knight (1997)
[5] Phenotype prediction from genome-wide association studies: application to smoking behaviors by Dankyu Yoon, Young J. Kim, and Taesung Park (2012)
[6] Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks by Bing Han, et. al. (2012)
[7] Gene Selection for Cancer Classification using Support Vector Machines by Guyon, et. al. (2002)
[8] Genetic Engineering and Cloning: Focus on Animal Biotechnology by Giassettia et. al.