Stat.585X: Draft of Project Plan

Lindsay Rutter

March 26, 2014

Project Topic

The project will focus on the construction and interactive visualization of phylogenetic representation of soybean varieties.

Project Motivation

The motivation of the project stems from my interest in bioinformatics and computational biology. Ideally, the product could be used by biologists, geneticists, and agronomists interested in studying how soybean varieties are related. By referencing the interactive visualization of the phylogenetic tree, these scientists may better understand genetic testing results - in this particular dataset, in terms of copy number variants, single nucleotide polymorphisms, protein content, and yield - and use that knowledge in future breeding.

Available Data

The available data consists of a data frame structure that contains 412 direct child-parent relationships between pairs of soybean varieties. These data were collected from friel trials, genetic studies, and United States Department of Agriculture (USDA) bulletins, and date as early as the first decade of the 1900s.

Data Collection and Processing

The data frame format of the soybean varieties has been represented as an interactive phylogenetic tree, in Shiny software produced by Susan VanderPlas:

http://gsoja.agron.iastate.edu:3838/Soybeans/

In the "Genealogy" tab of the Shiny application, the user may choose one or more varieties in the left-panel menu, and immediately view the phylogenetic tree representation of the selected varieties in the right-panel plot as per reactive programming.

However, currently, the right-panel plots are plotted independently for each soybean variety selected, showing a user-selected number of generations surrounding that variety, regardless of its relationship to any other varieties selected.

Instead, it may be useful for biologists to obtain one large plot in the right-panel that merges the selected soybean varieties, thereby determining not only whether or not there exists a relationship between the varieties, but also showing the relationship as a path in the graphical phylogetntic structure, with possibly the varieties (nodes) and the path (edges) between them highlighted for emphasis.

In total, there are about 230 unique soybean varieties present in the tree data frame. The package igraph might be used to determine any pathway relationships between selected varities:

http://cran.r-project.org/web/packages/igraph/index.html

Final Product

If the Shiny application is properly functioning, it will then be extensively commented in the format that will soon be taught in class. This may bring the application one step closer to being available to interested users via an R Package.