Statistical analysis: rice phenotype – biology project

Houda Aiboud Benchekroun and Thomas Roiseux (group 8)

2022-11-20

General information

Genetic data hold important information about living beings. These information can be used to understand and even predict the ability of a species, or an individual, to protect itself from a specific disease. Using this and statistical tools, we may be able to find some predispositions for individuals of a population to have or a not specific disease, and by making predictions, to find the most suitable gene that will grant the best protection against several problems.

To explore this, we will use a data set made with genetic information of rice. These have been extracted from the UK Bio-Bank. As rice is one of the basics in food in several countries, it production must be protected to avoid any starvation in these countries. That's why studying its genetic data to protect it is quite important nowadays.

To understand that, we are going to study phenotypic characters and explain them using genetic data coming from the genotype of this rice species, named *Oryza sativa*.

Variables available

Genomes

The data set we have is made of two different parts: we have a genome set and a phenotype set. The phenotype data set holds data about expressed characters of the rice, whereas the genome dataset holds all available alleles for rice genes. 413 accessions have been sampled in this data set, referenced in their genome. The genomes and its variables will be used as explanatory variables, as genes and alleles grant different visible characteristics to the rice. The 413 samples included in this data set reflect the 413 accessions from the total of sampling the rice species genomes, that have grown across the planet (Figure 1).

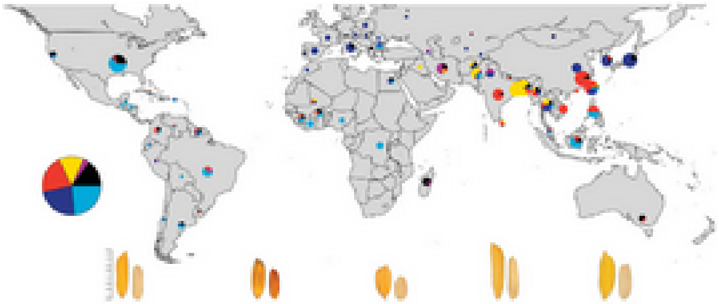

Figure 1: World map with all locations where the rice has grown

This map shows the distribution of the 413 species used to make the data set.

These variables, crossed with the genetics markers will provide explanations to the rice phenotype. They are all gathered in the `Xmat` matrix: in this matrix, each column is a genetic marker, and each row represents a specific accession of the rice. using this matrix, we know all the genetic data of each rice species. This will be important to understand and explain the phenotype.

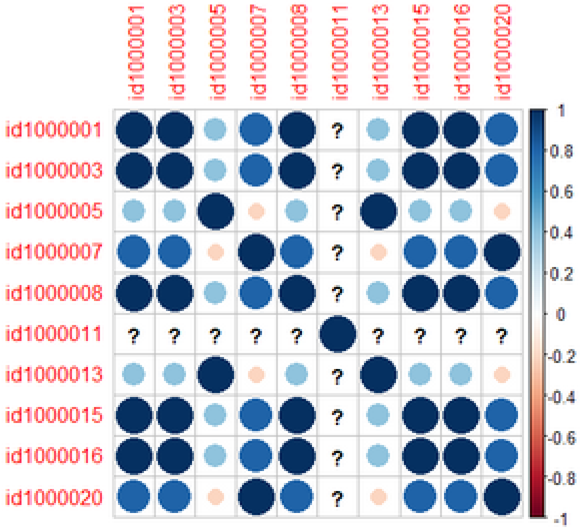

Figure 2: Correlation matrix for the first 10 explanatory variables

Using this plot of the correlation matrix on Figure 2, we can assume that we will select variables as, when using only ten variables, we can already notice strong correlation between some of them, like between `id1000001` and `id1000003`.

Phenotype

Phenotype holds all the visible and expressed characteristics of an individual, coming from its genetic information. This means that the phenotype directly depends on the genome of the individual, which is described and specified above.

In this data set, we are going to explain how the genetic information from the genomes can influence the protein content, hold in the `Protein.content` variable of the phenotype, and try to find which genes are mostly important to this characteristic.

This variable is quite important in determining the final quality of the rice: the more protein there is, the best the rice will be. Using classification methods, we will try to find the genes that contributes to a higher level of protein in the rice.