Shulei Yang

Naked Mole Rat Analysis

Statistical Consulting, Dr. Karen Kafadar

University of Virginia

Apr.27, 2020

1. Introduction

Naked mole rats fascinate scientists for so many reasons. They rarely get cancer, are resistant to some types of pain, have a long lifespan and can survive up to 18 minutes without oxygen. This report ultimately tries to answer the question: how female naked mole rats are different from male naked mole rats. And how breeders are different from nonbreeders. To answer this question, we will try to find which organs of a naked mole rat are most predictive of its sex and whether it is a breeder or not. This report consists of six sections. In 'Background', I will briefly introduce the background of this project and discuss the data set we used. Then in 'Method Overview', I will discuss the data cleaning process and introduce all the technics I used in my analysis. In the 'Results' section, I will present the results from my analysis and carefully explained all the results. Finally I will discuss the limitations of my analysis and draw my final conclusion.

2. Background

a. Eusocial Mammal

Eusociality, the highest level of organization of sociality, is defined by the following characteristics: cooperative brood care, overlapping generation within a colony of adults, and a division of labor into reproductive and non-reproductive group. Honeybees' sociality is a well-known example of eusociality. There were only two known mammal species which are eusocial. They are the naked mole rat and the Damaraland mole rat.

b. Naked mole rat

The naked mole rat is a burrowing rodent native to parts of East Africa. It has a highly unusual set of physical traits that allows it to thrive in a harsh underground environment. The tunnel systems built by naked mole rat can stretch up to three to five kilometers in cumulative length.

Naked mole rat's eusocial structure is similar to that found in bees. Only one female (the queen) and one to three males reproduce, while the rest of the members in the colony (despite male or female) function as workers. Workers are sterile, the non-reproducing females appear to be reproductively suppressed, meaning their ovaries do not fully mature and do not have the same levels of certain hormones as the reproducing females.

Naked mole rat is a special mammal worth study for a lot of reasons. First of all, the naked mole rat holds the records for the longest living rodent, it can live up to 31 years. Also, the mortality rate of this rat does not increase with age and it is fertile throughout its lifespan. Finally, naked mole rat has a high resistance to cancer and it is pain insensitive.

c. Dataset

The dataset consists the RNA information for 100 observations. Among the 100 observations, there are 25 breeder females, 25 non-breeder females, 25 breeder males and 25 non-breeder males. In total there are 166353 . The 166353 variables present different RNA extracted from ten different tissues (heart, skin, liver, kidney, cerebellum, hypothalamus, pituitary, thyroid, adrenal, gonads)

3. Method Overview

a. Data Cleaning

i. Group RNA sequence by 'tissue'

166353 variables were too large for our analysis, especially considered the fact that we only have 100 observations. Therefore, I created a new dataset, where all the RNA variables are grouped by 'tissue'. There are in total 10 tissues, and the value of each tissues are the sum of all RNA included. Therefore, we transform the question into which parts (tissues) of naked mole rat are most predictive of its sex and whether it is breeder to not.

ii. Use Logarithm to scale the dataset

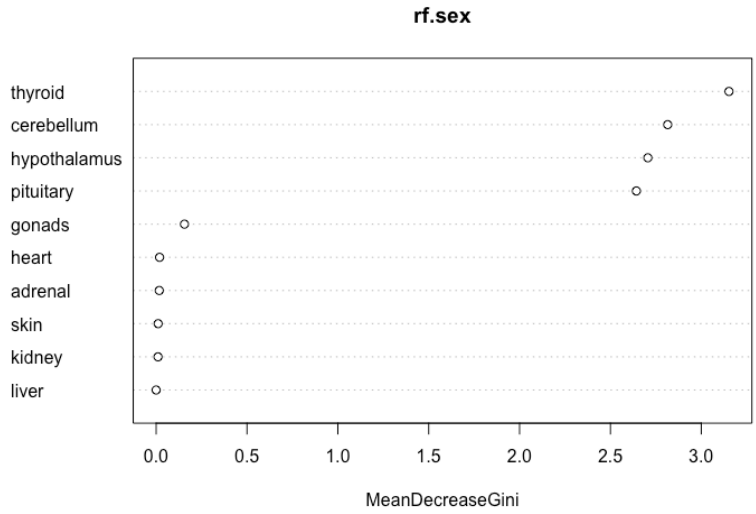
This dataset has a large span in values. Some data can be as large as '165,234', while some data is only as small as '982'. To prevent the skewness towards large values, I applied logarithm to the whole dataset. By doing this we can manage to bring the values to the same magnitude and improve the accuracy of our analysis.

b. Random Forest

Random Forest is an extension of the decision tree method. A decision tree method is a flowchart-like structure in which each node represents a 'condition' on an attribute (e.g. whether a coin flip comes up a head or a tail), each branch represents the outcome of the 'condition'. The decision tree method can help us identify the important variables that lead to the final result. However, the decision tree method only produces a single tree, therefore, its result might be limited. Therefore, in this project I used the random forest method. The overall idea of random forest is to aggregate a large number of decision trees and come up with a more robust decision. Random forest 'learns' individual trees with some random perturbation, and then 'average' these trees. Random forest method can tell us which variables are most influential for the final outcomes (what tissues cause the difference in sex and what tissues are most predictive of whether an observation is breeder or nonbreeder).

4. Results

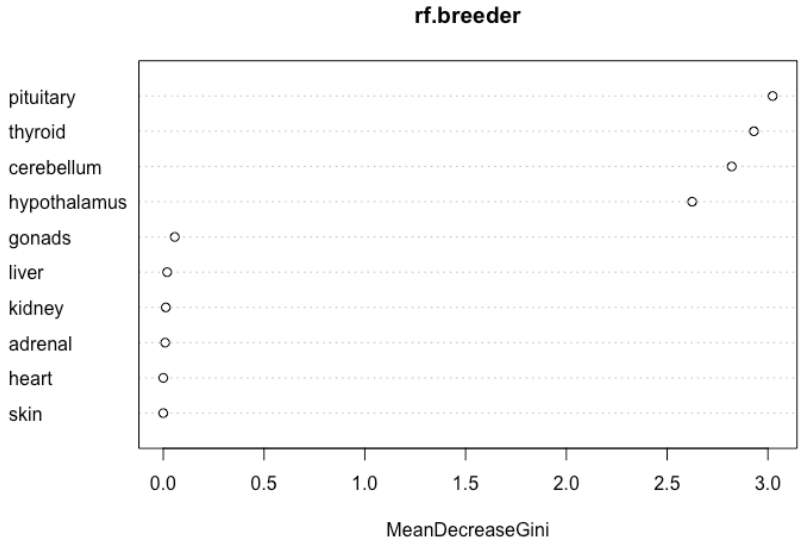
a. Differences between male and female



The above plot shows the importance of each variables in a random forest model. The variables appear at the top of the plot with a large 'MeanDecreaseGini' value reduces more impurity in the random forest. This indicates that the variables at the top of the plot are more important in deciding whether an observation is male or female.

From this plot we can see that RNA information from 'thyroid', 'cerebellum', 'hypothalamus' and 'pituitary' are more important in discriminating whether an observation is female or male. However, the 'MeanDecreaseGini' of 'heart', 'adrenal', 'skin', 'kidney', and 'liver' are all zero. This means that the RNA information in these variables are roughly the same for female and male naked mole rat. We should also notice that surprisingly, the RNA information from 'gonads' is not that different for male and female according to this random forest model.

b. Differences between breeder and non-breeder



From this plot we can see that RNA information from ‘pituitary’, ‘thyroid’, ‘cerebellum’ and ‘hypothalamus’ are more important in discriminating whether an observation is breeder or non-breeder. Still, the ‘MeanDecreaseGini’ of ‘heart’, ‘adrenal’, ‘skin’, ‘kidney’, and ‘liver’ are all zero, means that the RNA information from these variables are roughly the same for breeder and nonbreeder naked mole rat.

5. Limitations

We should notice that there is a potential limitation of my results. Result from random forest model might suffer from overfitting when the sample size is small. In this project, the total sample size is only 100, therefore the results might be only statistically meaningful but not statistically meaningful. However, since this is an ongoing experiment, your team are still continuously collecting more data. I believe this random forest model will be more accurate when more data can be used to train this model.

6. Conclusion

Combing the results above, I found that the RNA information from ‘pituitary’, ‘thyroid’, ‘cerebellum’ and ‘hypothalamus’ determines both the sex and whether an observation is breeder or not. However, the RNA information from ‘heart’, ‘adrenal’, ‘skin’, ‘kidney’, and ‘liver’ are similar for all naked mole rat.