Network Tour of Data Science project

Genetically determined susceptibility to malaria

Valérian Rey, Rayane Laraki, Maxence Jouve, Artur Szałata

Supervisor: Benjamin Ricaud Lausanne, 10.01.2020

1 Objective

It is a well known fact that DNA determines peoples' immune responses. In general genes determine gene expressions which in turn influence the organisms phenotype.

The aim of this project is to establish concrete links between genes' expression in particular tissues and susceptibility to malaria in mammals, in particular mice. We have only partial information about genes' expression, so we try to infer the missing data. Knowledge about the influence of gene expression and disease immunity could be, after extensive research, extrapolated to humans and those more susceptible could be advised to change lifestyle and have more frequent screenings to reduce disease risk.

2 Data description

We use a dataset of gene expressions and information on susceptibility to malaria of mice of BXD strains. It is a subset of the open dataset available at the genenetwork website and the link to it is in the README file of our repository. The data we use consists of:

- Information on malaria susceptibility comes from "Phenotype.txt" marked with "X4233" PhenoID which corresponds to

  "Malaria susceptibility, murine Plasmodium yoelli 17X-lethal (PY17X-L) [0=100% resistant, 1= 100% susceptible]"

  phenotype.
- Gene expression from all the tissues stored in separate text files in the "expression data" directory.

3 Summary

3.1 Major Steps

1. We will establish a baseline for the graph based approach, that is try to predict the immune response directly from the genes' expression using simple ridge regression.
2. Build a coexpression graph: distance metric is based on the euclidean distance between the expression of genes in different mice.
3. We infer gene expression where it is missing by applying Tikhonov regularization for each mouse on the coexpression graph.
4. We try to predict the immune response of the mice using the baseline approach from point 1 on the data with inferred expression and compare the results with baseline. Improvement in the prediction means that we somewhat correctly propagated the phenotype.

3.2 Tools

We base our work on two algorithms:

- Ridge regression for predicting malaria susceptibility given gene expression in a mouse. Link between malaria and DNA has been established previously [1]. We use only such a simple model as we have very little data points, 57 mice, given many features, over 1M gene expressions per mouse (with plenty of missing data).

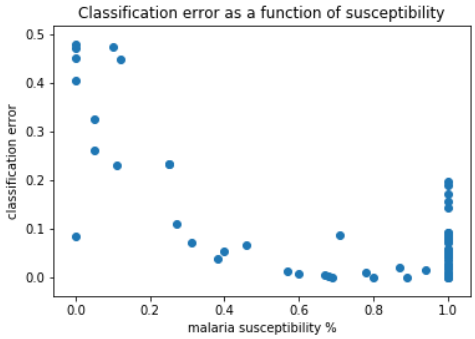

Figure 1: Baseline model using 1.2M features

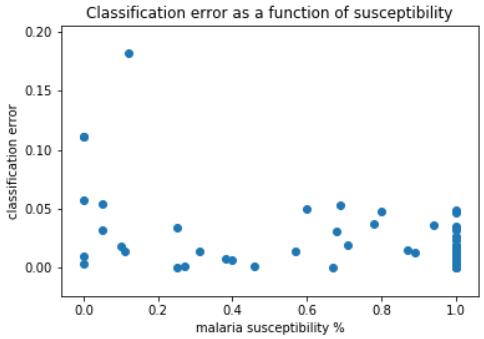

Figure 2: Baseline model using 800 features. Note y scale.

- Tikhonov regularization on coexpression graph, knowing that there are genes that influence expression of other genes [2]. Such relation makes it sensible to smoothen the signal on the coexpression graph to infer missing values for each mouse.

4 Approach

Detailed explanation of our work.

4.1 Data Processing

We have expression data of same genes in multiple tissues. We decide to treat those as separate gene expressions and later call each such SNP expression in a tissue (SNP, tissue pair) an SNP expression. We noted that the values of expression appear to follow normal distribution and so we have standardized them for each SNP expression: we subtract the mean and divide by standard deviation. To avoid basing our results on very limited data, we decide to drop all the SNPs for which we have data in less than 5 mice. Each mouse has on average 48.7% of the expression data missing. For each SNP expression we fill this missing data with the mean to facilitate prediction.

4.2 Prediction Baseline

As we have 1.2M features and only 57 data points, we are limited to using very simple machine learning models. We decide to use ridge regression. Initially we wanted to make sure that prediction is indeed feasible given the data and so we created 20 random train-test splits with a test set making up 30% and looked at the percentage of positive R^2 scores of the ridge regression model on them. Recall that values of R^2 score range from $-\infty$ to 1, where 1 means perfect prediction and score 0 is attained by a constant model predicting the expected value of y disregarding the features. Thus a majority of positive scores gives an indication that model is successfully using the features to predict the outcome. We found out that R^2 scores are mostly positive, so we decided to use this approach on the malaria susceptibility. Next, we look at the squared error our model makes, using cross-validation as we have very limited dataset. We note that mean squared error is 0.114 where $y \in [0, 1]$. We see that proportion of mice with high susceptibility is bigger than those with immunity, so we want to make sure that the model is not only good at predicting for one of the groups.

From figure 1 it is clear that the model performs better for mice with high susceptibility, having 0.066 MSE for completely susceptible mice compared with 0.164 for the rest.

We note that for most of the features ridge regression uses very small weights, meaning that they are nearly ignored. For this reason we decide to reduce the number of features used. As the way to find most relevant ones we pick those that were assigned weights with highest absolute value and those that have highest absolute value of spearman correlation with the malaria susceptibility. We noted lowest mean errors when using 50 000 features, but to speed up our experiments we decide to use only 800 features total: half using ridge regression weights as indicators and half using spearman correlation. Such a combination performed better than using a single criterion.

A model with only 800 features has 0.024 MSE compared to 0.114 of the one that uses all the 1.2M. We can see in 2 that it improves in general, not promoting only some of the mice.

In the end we decide to see predictive power of 10 features and surprisingly we got MSE as low as 0.068. Among the selected features there are 3 expressions in bone femur and 3 in brain INIA , so it seems that expressions in those

tissues are important indicators of malaria susceptibility. See "baseline_phenotype_predictor.ipynb" in our repository for the precise SNPs.

4.3 Graph Construction and Exploration

We built a co-expression graph between SNPs. Each node corresponds to a SNP in a given tissue (example: Tpp2_ILM3850093 in Femur). The distance between two SNP, X and Y is computed as follow:

- we first obtain the vectors (u and v) corresponding to the expression value of all strains for SNP X and Y.
- we then compute the number of common strains for these two vectors u and v, call it n.
- we then compute the euclidian distance e between the non-NaN values of u and v.
- we obtain the distance d between SNP X and Y by computing $d = \frac{e}{n}$ if $n \geq 10$ otherwise we have $d = n$.

Now that we have the distance matrix, we applied a RBF (Radial Basis Function) kernel with parameters σ (width of the kernel) and ϵ (threshold value). We initialized σ as the median L_2 distance between data points and then tuned both σ and ϵ to obtain a sparse matrix with dominating connected components. Finally we decided to keep the biggest connected component as the the other ones were containing very few nodes each.

Thus we obtained a co-expression graph containing 696 nodes and 15254 edges. Here is the degree distribution:

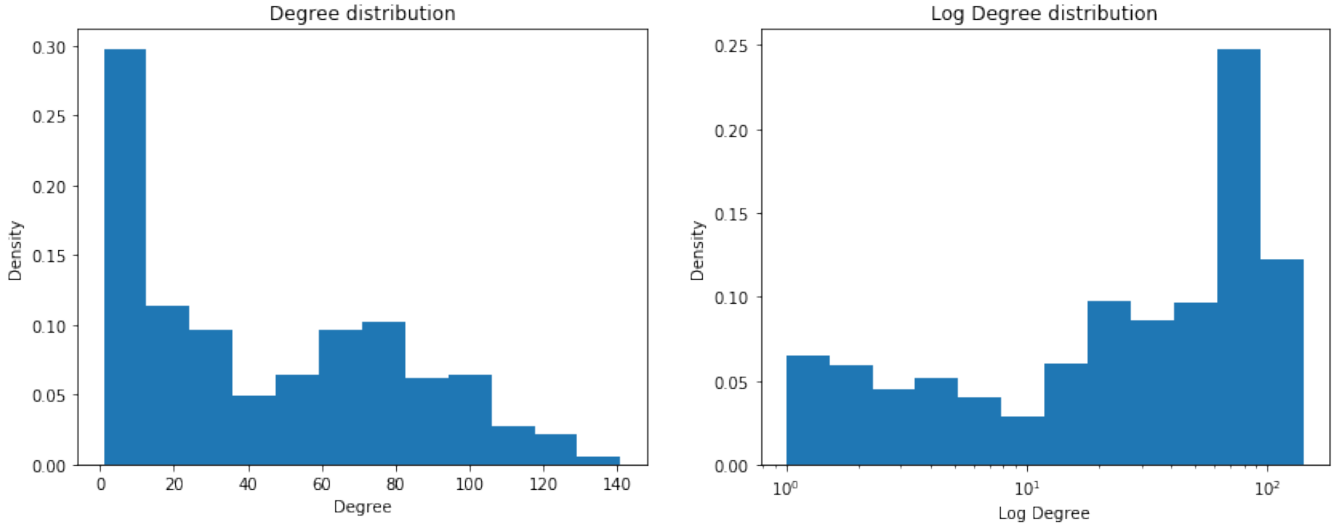

Figure 3: Degree Distribution of the Co-Expression Graph

As we can observe, the degree distribution is not heavy-tailed. That means that our graph does not have big hubs. The average degree is 43.8 and the maximum degree is 141 which is less than 5 times the average degree. Moreover the diameter of the graph is 16 which is a lot given the total number of nodes. Finally, the average clustering coefficient is 0.58. It means that overall nodes are highly connected between each others but we do not have hubs connecting most of the nodes in the graph (thus leading to an important radius and a high clustering coefficient).

4.4 Graph Visualization

In order to visualized the obtained graph, we used Gephi. We applied the Fruchterman-Reingold spatialization algorithm and we colored the edges depending on their weight (the larger their weight, the brighter they are). We did not display the nodes to keep everything simple and clean.

4.5 Imputation Using Co-Expression Graph

We assume smoothness of the expression signal on the created graph as we constructed it to have such property: the nodes are close if the expression is similar for multiple mice. With this assumption, to infer missing expression values for each mouse we put its expression as a signal on the coexpression graph filling missing values with mean

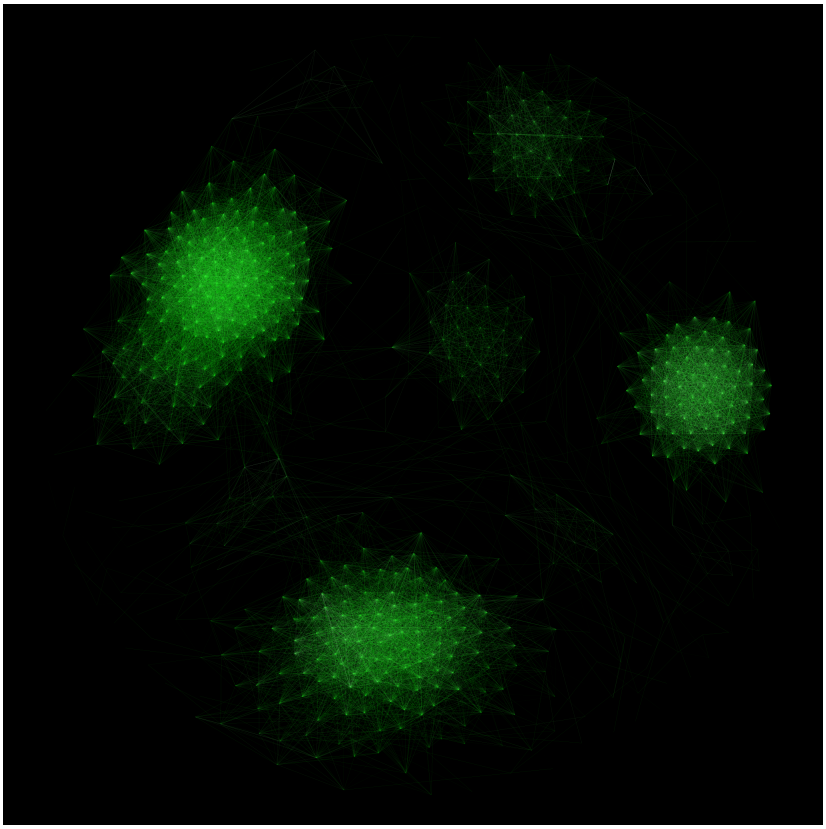

Figure 4: Co-expression Graph visulized with Fruchterman-Reingold algorithm

(as previously) and apply Tikhonov regularization. Such regularized signal, scaled appropriately (without scaling influence is small), is used as the inferred value of the missing expressions. Resulting graph for mouse 50 can be seen in figure 6. Recall the optimization problem of Tikhonov regularization:

$$\tilde{x} = argmin_{x \in R^N} \|Ax - y\|_2^2 + R_{tk}(x; G)$$

$$R_{tk}(x; G) = \alpha \|Sx\|_2^2$$

where A is the adjacency matrix and S is the incidence matrix.

The ridge regression model on the expression data with inferred missing values outperforms the original one by more than 7% in MSE, namely it has 0.0228 MSE.

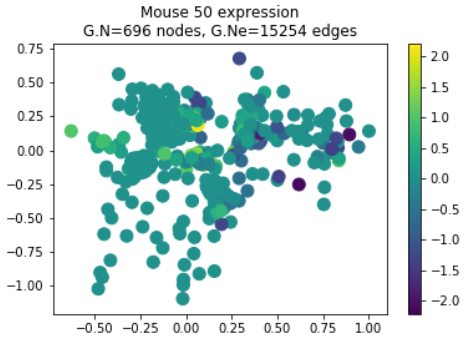

Figure 5: Expression with missing values set to 0

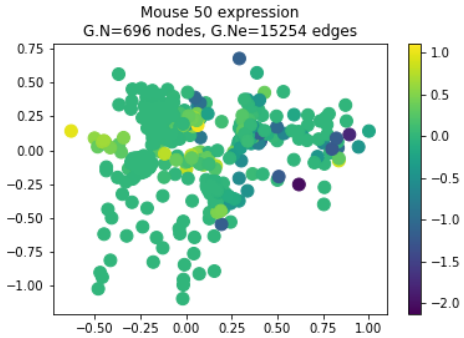

Figure 6: Expression after Tikhonov regularization

5 Possible drawbacks

- The dataset has missing fields which might make the analysis and results less relevant.
- We have very limited number of data points - only 57 mice with given malaria susceptibility. It is not desirable for machine learning techniques.

6 Future work

- We have notes best prediction scores when using 50 000 SNPs' expression data and experiments with such amount would certainly improve the final score. The only reason for using many less, a total of 800, were limited computational resources at our disposal.
- It is widely known that genes not only coexpress, but often also inhibit expression of other genes. One could build a graph of inhibition and use it together with coexpression graph to better infer missing data.
- We have limited ourselves to malaria susceptibility, but using nearly identical scheme, with different hyperparameters, one could likely perform similar prediction for other phenotype and identify the genes that influence it most.
- Note that our approach uses many hyperparameters such as: number of SNP expressions under consideration, epsilon and sigma in the RBF kernel used for constructing the adjacency matrix, gamma in the Tikhonov filter and many more. One could certainly find somewhat better combination of those given more time.

References

[1] Adel Driss, Jacqueline M. Hibbert, Nana O. Wilson, Shareen A. Iqbal, Thomas V. Adamkiewicz, and Jonathan K. Stiles. Genetic polymorphisms linked to susceptibility to malaria. 10(1):271.

[2] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. 19(4):575–592.