Genetically Determined Susceptibility to Malaria

Valérian Rey, Rayane Laraki, Maxence Jouve, Artur Szałata

École Polytechnique Fédérale de Lausanne (EPFL)

EE-558 Network Tour of Data Science

January 2020

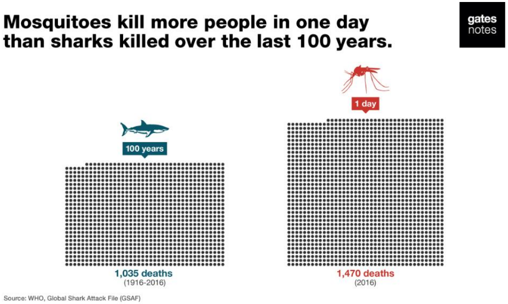
Outlines

1. Introduction
   - Problem
   - Dataset
   - Approach
2. Our approach
   - Baseline
   - Graph construction
   - Graph analysis
   - Imputation
3. Future work
4. Conclusion

Introduction

Problem

- Predict immune response given only genetic information
- Determine most relevant causes of susceptibility or immunity



Mosquitoes kill more people in one day than sharks killed over the last 100 years.

Source: WHO, Global Shark Attack File (GSAF)

Dataset

- genes' expressions in tissues and phenotype of BXD strains. A subset of the open dataset available at the genenetwork website[1].
- 57 strains with given malaria susceptibility score
- 1.2M genes' expressions per mouse with roughly 50% missing

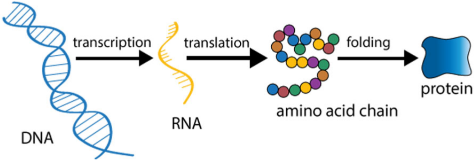


Figure: Gene expression[2]

[1] http://www.genenetwork.org

[2] image from www.researchgate.net

Approach

- Use only genes' expressions as features
- Establish a baseline using ridge regression with cross validation
- Pick relevant subset of genes' expressions data
- Build a coexpression graph with genes' expressions as nodes and apply Tikhonov regularization to infer the missing data
- Apply ridge regression on data with inferred expressions

Our approach

Baseline

- Select features: gene-tissue expression
- Explore the data, standardize and fill missing values
- Evaluate ridge regression using cross validation. MSE 0.114 (33%).

Most important bits

Select a subset of features using ridge regression weights and spearman correlation with malaria susceptibility

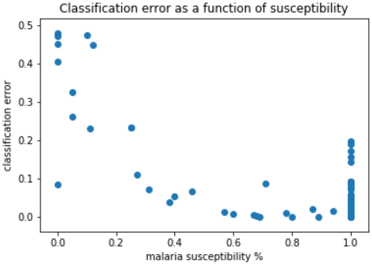


Figure: Baseline model using 1.2M features

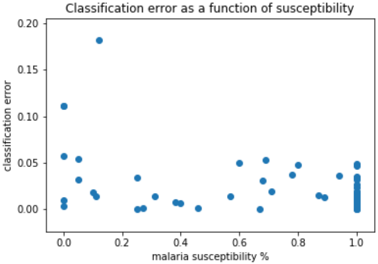


Figure: Baseline model using 800 features

Most relevant genes

10 most relevant genes
Serpina1-rat_ILM106590035_**Bone_Femur**
Cntnap2_ILM100380601_**Bone_Femur**
1700072I22Rik_ILM100380731_**Bone_Femur**
D19Ertd678e_1441578_at_B_**Brain_INIA**
2900041M22Rik_1444801_at_B_**Brain_INIA**
Cdc40_1445348_at_B_**Brain_INIA**
Gm16000_TC0300002214.mm.1_ScWAT_HFD
2510015N06Rik_1441597_at_Kidney_Male
_TC1700002137.mm.1_ScWAT_CD
Rtl1_10398346_Adrenal_Female

# Features used	MSE
1.2M (all)	0.114 (33%)
56k	**0.012 (11%)**
800	0.024 (15%)
10	0.068 (26%)

Graph construction

Goal: Build a co-expression graph between genes' expression. Each node corresponds to a gene in a given tissue.

- Example: Tpp2_ILM3850093 in Femur

Computing the distance matrix

Here is how we computed the distance between two genes (nodes) X and Y in the graph:

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

From the distance matrix to the final Graph

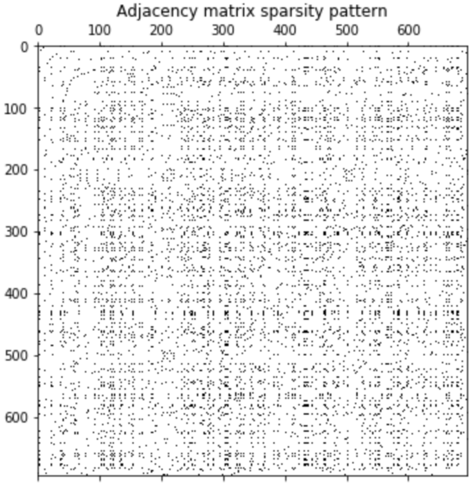
1. Apply a RBF (Radial Basis Function) kernel with parameters σ (width of the kernel) and ϵ (threshold value) on the distance matrix.
2. Initialize σ as the median L_2 distance between data points and then tuned both σ and ϵ to obtain a sparse matrix with dominating connected components.
3. Keep the biggest connected component as the the other ones were containing very few nodes each.

We obtained a co-expression graph containing 696 nodes and 15254 edges.

Graph analysis

- Some properties of the graph
- Some properties of the nodes
- A visualization of the network

Properties of the graph



- The graph has 1 connected component
- The diameter is 16
- The average clustering coefficient is 0.5832621152157119

Degree distribution

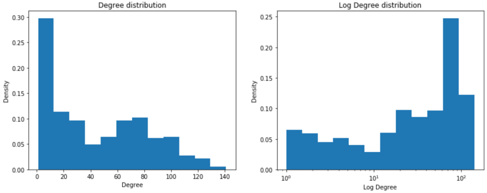


Figure: Degree distribution

- The distribution is a bit heavy-tailed. That means that our graph does have some hubs, but not very big (Average degree is 43.8; Maximum degree is 141).

Graph visualization

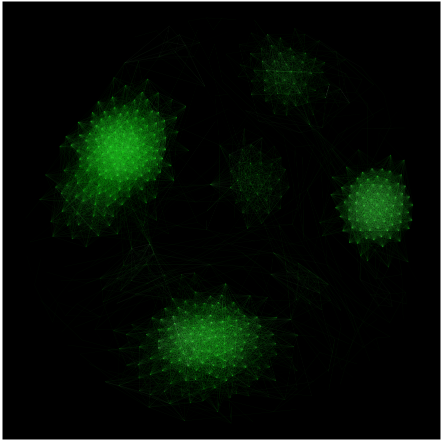


Figure: Fruchterman-Reingold visualization

- Fruchterman-Reingold visualization of the graph.
- Edges with heavier weights are brighter.
- Some clusters seem to appear.
- Interactive visualization

Imputation

- Signal is the genes' expression for each mouse in turn
- Smoothness assumption on the coexpression the graph
- Tikhonov regularization to infer missing values

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

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

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

Imputation

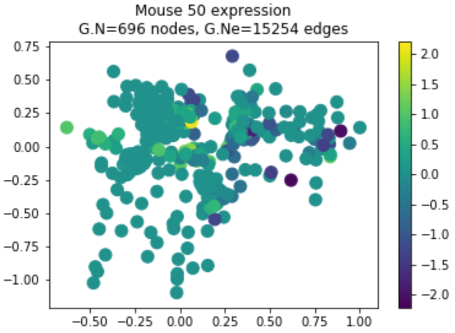


Figure: Expression with missing values set to 0

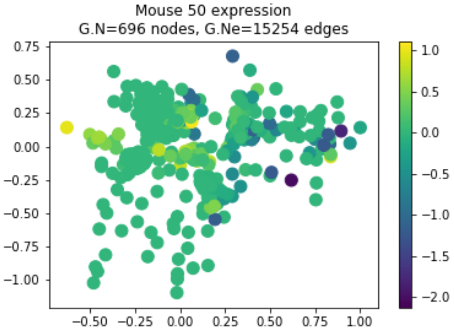


Figure: Expression after Tikhonov regularization

Results

Missing value policy	MSE
Fill with mean	0.02445 (15.64%)
Tikhonov regularization	**0.02269 (15.06%)**

Future work

What next?

- Use more genes' expressions
- Inhibition graph
- Predict other phenotypes
- Hyperparameter tuning
- Regularization on mice strains graph for phenotype prediction

Conclusion

Conclusion

- Pros
  - Effective inference of missing data
  - Very high accuracy in phenotype prediction
  - Can identify most significant genes (even 10 say much!)
- Cons
  - Missing fields in the dataset
  - Only 57 strains with given phenotype

Thank you for your attention