BXD Mice Genetics - Project Summary

Gianni Giusto, Yann Mentha, Raphaël Reis Nunes and Lucas Zweili

1. Introduction and motivation

What if we could predict a coming genetic disease from a simple blood test? Can we diagnose a patient which is at high risk of developing a cancer or suffering from heart failure in the next 2, 5 or even 10 years? Often, a disease can be seen as a combination of several phenotypes which lead to body failure. In such cases, if we can predict the latter phenotypes, we could eventually be able to diagnose such a disease [1].

As a first step towards this personalized medicine ultimate goal, we will investigate on basic phenotype prediction using the genomic dataset[1] from the BXD mice family[2]. This dataset contains information about genetics, protein expression and phenotypes for each mouse strain[3]. But due to many reasons, the data acquisition led to lots of missing values. In order to produce a sound prediction, we aim at estimating missing information using graph signal processing (GSP) methods, taking advantage of the network structure of the data, and compare it with a baseline prediction without these value estimations.

2. Data acquisition

The data consists mainly of three type of files: a *genetic* file where a binary value tells from which parent a certain position in the genome is inherited, a *phenotype* file summarizing the expression level of a phenotype for each strain and several *expression* files where multiomic clinical and molecular phenotypes are described.

Each mouse strain represents a node of a network. They are linked together based on a similarity measure (see after for more details). Each node is associated to several attributes (features) such as the expression level of a certain protein for a given phenotype.

3. Data exploration

Intuitively, it seems reasonable to build the main graph based on genetic information: mice that have a similar genetic material will be linked stronger than mice with low genetic resemblance. Note that all mice must be connected in order to perform signal processing analysis. Preliminary analysis have led to the graph characterized in Table 1.

4. Exploitation

The current dataset is a result of experiments pooled together from several laboratory around the world. Thus, it is not surprising to find a significant number of missing values (namely NaN), see Fig. 1. The challenge will be to transform the node features in the Fourier domain and find an appropriate filter to exploit the network structure of the mice genetics (see Fig. 2 for high variance molecular expression GFT). We expect this regularization to spread the information through the nodes, replacing missing values by relevant ones according to the genetic similarity. We will first implement this approach on a limited set of features (10 to 15) and scale up thereafter.

As a second step, we will build a classifier that aims to predict node phenotype (label) based on molecular expression and linkage properties of the network. We expect a result suggesting that missing values inference from GSP method brings relevant information to our classifier model.

[1] Original dataset: http://www.genenetwork.org || sub-sample we used see: https://drive.switch.ch/index.php/s/mtQ2F0dYc7dHOtQ

[2] The BXD family is a set of mice which were derived by crossing two genetically specific mice (C57BL/6J (B6) and DBA/2J (D2)). Each offspring is uniquely defined in genetic terms.

[3] A strain is a mouse which is uniquely defined in the BXD family from the genetics point of view.

Tables and Figures

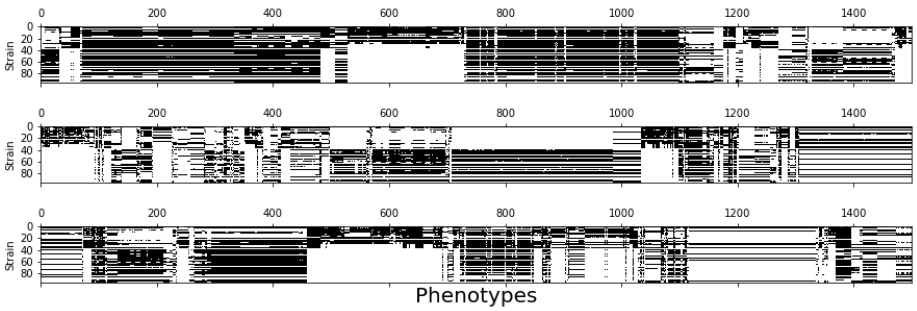
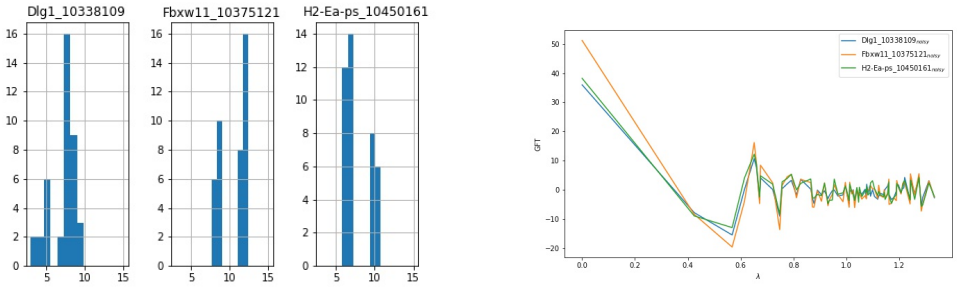


Figure 1: **Visualization of missing values.** For each combination of mouse strain and phenotype, missing values are represented by white areas whereas black spots depict known information.



(a) (b)

Figure 2: (a) **Distribution of high variance molecular expression.** (b) **Graph Fourier transform for the molecular expression data.**

Table 1: Genetic graph characteristics.

Genetics graph	
Number of Nodes	97
Number of Edges	2226
Graph density	47.81%
Average Degree	45.9
Number of Connected Components	1
Diameter of the network	3
Average Clustering Coefficient	0.54

References

[1] H. Li, X. Wang, D. Rukina, Q. Huang, T. Lin, V. Sorrentino, H. Zhang, M. B. Sleiman, D. Arends, A. McDaid, P. Luan, N. Ziari, L. A. Velázquez-Villegas, K. Gariani, Z. Kutalik, K. Schoonjans, R. A. Radcliffe, P. Prins, S. Morgenthaler, R. W. Williams, and J. Auwerx, "An integrated systems genetics and omics toolkit to probe gene function," *Cell Systems*, vol. 6, no. 1, pp. 90 – 102.e4, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405471217304866