Matthieu Simeoni
Matthieu Gendre
Guéguen Louis-Maël

Autoencoding diseases

Introduction

Today, with the development of the WGS, many studies such as the GWAS have emerged. These studies have revealed numerous genetic variants associated with certain phenotypic traits.

We can group the different GWAS together to build a network of human diseases, which is a tool to visualize the possible relationships between different phenotypes and on which genes these relationships rely on.
More specifically, the human disease network allows us, for example, to predict possible diseases related to genes, and observe whether certain diseases have a common genetic origin or not.

One important thing to know is that until now, human disease networks have been constructed according to the literature. More precisely, a link between two phenotypic traits is established if these two trait terms coexist sufficiently in the literature. From a graphical point of view, the more these links are present in the same literature, the closer two nodes will be graphically. Thus, the main goal of this project is to establish a link between the two phenotypic traits using genotypic data instead.

To do so, we built networks from the GWAS data set using a neural network and not through literature searches.

More specifically, the neuronal network will be an autoencoder: it gives rise to an output that is similar to the input. We can consider autoencoders as an unsupervised learning technique, which means it looks for previously undetected patterns in the data with no pre-existing labels.

Methods

We used GWAS data for our project and manipulated it using Pascal and Python.

We first collect raw GWAS statistical summaries to build our dataset. Then, we use Pascal to convert SNP information to gene-level information. More precisely, we obtain the likelihood of how related each gene is to the GWAS phenotype.

Then, all processed GWAS is merged into a matrix to perform various operations using Python.

Pre-processing steps were performed on the matrix in question. In particular, missing values are replaced by uniformly distributed random values, a Bonferroni correction of the thresholds is applied, transformations (-log10) were tested in order to keep only relevant information, and genes were filtered to keep only the ones that are significant for at least one disease.

We were then able to feed the data into the autoencoder built essentially from Keras and Sklearn.

Finally, the dimensional space of the autoencoder middle layer was extracted to construct a adjacency matrix that was used to plot the networks.

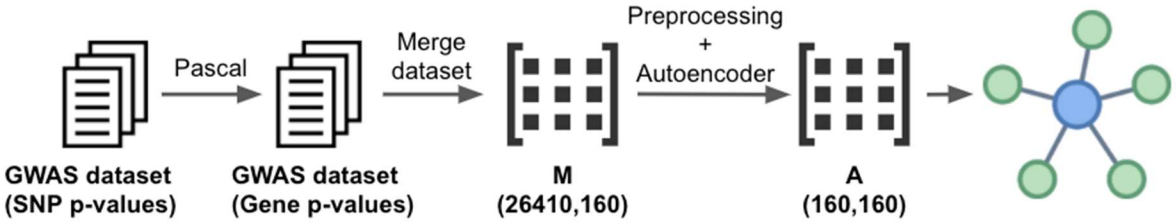

Figure 1 : Overview of the different steps carried out during the project. The extraction of GWAS data, the use of the Pascal algorithm, the manipulation of data, and the use of the autoencoder.

GWAS collection

A collection of 180 GWAS of data was already available to us beforehand, coming from Dream Challenge. But some categories were very present in the data, notably, the anthropometric category, and this category is not a disease.

In order to avoid this bias, we added 17 GWAS of several disease categories from the GWAS catalogue.

In addition, we removed certain types of redundant traits that were present in the original dataset, such as BMI.

Finally, we obtained a dataset of 160 GWAS files with an almost uniform distribution of the phenotypes.

PAthway SCoring Algorithm: PASCAL

PASCAL aggregates SNP statistics into gene statistics. It requires either the original genetic information of each person, either an external dataset used as a reference. This reference is the European subset of the 1KG project. PASCAL performs through the following steps:

- SNPs values are selected within a window around the gene sequence, with a MAF (minor allele frequency) of 5%. The MAF threshold ensures the SNP actually holds information about phenotypic variation. If the SNP does not vary enough along with phenotypes, it is considered as irrelevant for the math following.
- A theoretical distribution of the scores (derived from SNP_{GWAS}-SNP_{1KG} correlations) is calculated for each SNP, then the observed distribution is calculated as well.
- Both are compared and it gives rise to a p-value for the gene.
- These steps are performed for each gene for each phenotype.

The result is a file containing genes and their corresponding p-values, for each phenotype. These p-values are the likelihood that the gene is related to the phenotype; how related is the gene to the disease, per say.

Matthieu Simeoni
Matthieu Gendre
Guéguen Louis-Maël

Autoencoder and adjacency matrix

Concerning the application of the autoencoder, we have set up two autoencoders, one with a linear model and one with a sigmoidal model. But we decided to present only the sigmoidal model in relation to the results obtained.

The dimensional space was extracted from the middle layer of the autoencoder and we computed a Euclidean distance matrix of the size (160,160).

Then, from the Euclidean distance matrix, we calculated an adjacency matrix which is a matrix scaled from 0 to 1, that gives the strength with which the nodes are connected to each other. Thus, the more two diseases are connected, the closer the value will be to 1.

Results

From the adjacency matrix, a plot has been plotted using Networkx with firstly a cutoff at 0.8. The network obtained has 91 GWAS thus 69 GWAS have been removed with the cutoff. Structurally, a blob is present which it could be due to a possible high betweenness centrality of the nodes (figure 2). Therefore, to precisely analyze the blob and keep only the strongest connections, a second cutoff was made at 0.98.

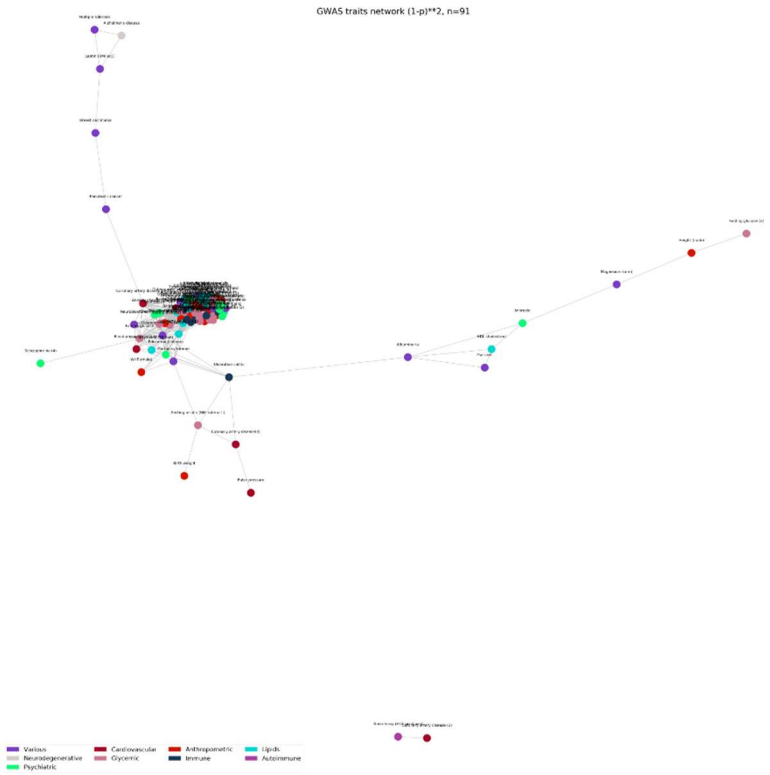

Figure 2 : GWAS trait networks composed of 91 GWAS. Transformation applied on the data is $(1-p)^2$. Cutoff applied to the adjacency matrix is 0.98.

In this case, the network has 54 nodes and we focus on few clusters (figure 3).

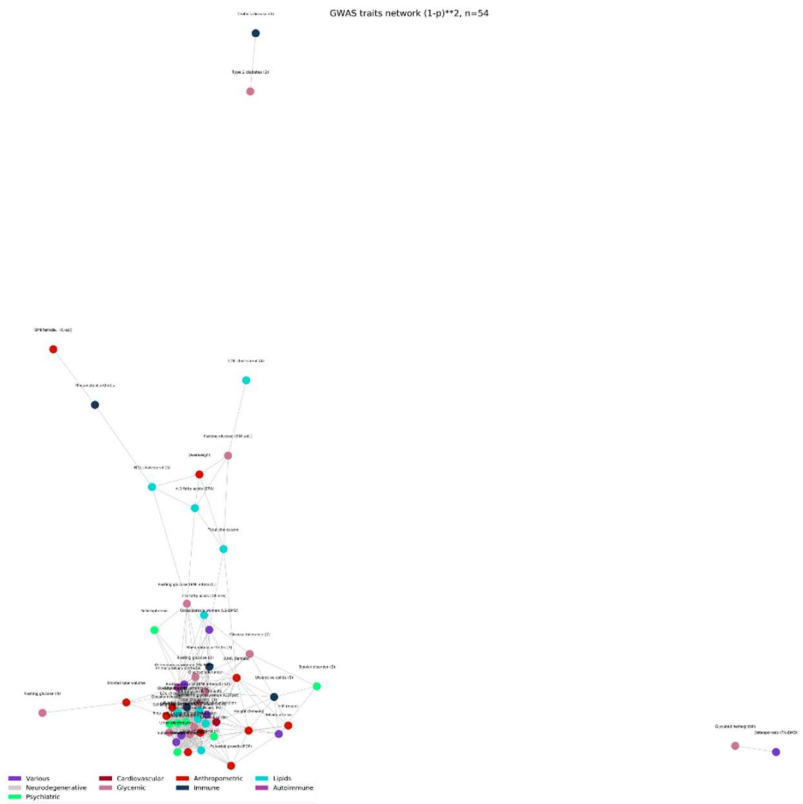

Figure 3 : GWAS trait networks composed of 54 GWAS. Transformation applied on the data is (1-p.value)². Cutoff applied to the adjacency matrix is 0.98.

Crohn's disease / Type 2 diabetes

In the literature, a link between these two diseases have been demonstrated but by not proving that the link is genetic. A patient has a significantly risk developing diabetes if he has a Crohn's (Kang *et al.* 2019).

Ulcerative colitis /HIP (male) /Biliary atresia /Height (female)

This cluster seems to be a new one because there is not reported links in the literature between all these traits.

Fasting glucose / Overweight / n-3 fatty acids / Total cholesterol

Concerning this cluster, each relationship has been described in the literature, but no paper has put forward a possible genetic relationship between these traits.[1, 2]

[1] *Campbell, S.C. and Bello, N.T. (2016). Omega-3 Fatty Acids and Obesity. Journal of Food and Nutritional Disorders, [online] 2012.*

[2] *Censin, J.C., Peters, S.A.E., Bovijn, J., Ferreira, T., Pulit, S.L., Mägi, R., Mahajan, A., Holmes, M.V. and Lindgren, C.M. (2019). Causal relationships between obesity and the leading causes of death in women and men. PLOS Genetics, 15(10), p.e1008405*

Discussion and limitations

Several points for improvement can be highlighted.

Especially about the autoencoding part: we could try different distance methods such as Manhattan, or Minkowksi. Moreover, we could also try different activation functions (elu, swish, ...).

About the clusters, a more precise inspection of the biological meaning of the clusters could be done. Moreover, we could use mathematical tools to analyze them (graph theory).

References

Campbell, S.C. and Bello, N.T. (2016). Omega-3 Fatty Acids and Obesity. Journal of Food and Nutritional Disorders, [online] 2012.

Censin, J.C., Peters, S.A.E., Bovijn, J., Ferreira, T., Pulit, S.L., Mägi, R., Mahajan, A., Holmes, M.V. and Lindgren, C.M. (2019). Causal relationships between obesity and the leading causes of death in women and men. PLOS Genetics, 15(10), p.e1008405.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. Proceedings of the National Academy of Sciences, 104(21), 8685–8690.

Jess, T., Jensen, B.W., Andersson, M., Villumsen, M. and Allin, K.H. (2020). Inflammatory Bowel Diseases Increase Risk of Type 2 Diabetes in a Nationwide Cohort Study. Clinical Gastroenterology and Hepatology, [online] 18(4), pp.881–888.e1.

Kang, E.A., Han, K., Chun, J., Soh, H., Park, S., Im, J.P. and Kim, J.S. (2019). Increased Risk of Diabetes in Inflammatory Bowel Disease Patients: A Nationwide Population-Based Study in Korea. Journal of Clinical Medicine, [online] 8(3).

Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput Biol 12(1): e1004714. doi:10.1371/journal.pcbi.1004714