Deriving Disease Modules from the Compressed Transcriptional Space Embedded in a Deep Autoencoder

Sanjiv K. Dwivedi [1] **Andreas Tjärnberg** [1 2 3] **Jesper Tegnér** [4 5 6] **Mika Gustafsson** [1]

Abstract

Disease modules in molecular interaction maps have been useful for characterizing diseases. Yet biological networks, that commonly define such modules are incomplete and biased toward some well-studied disease genes. Here we ask whether disease-relevant modules of genes can be discovered without prior knowledge of a biological network, instead training a deep autoencoder from large transcriptional data. We hypothesize that modules could be discovered within the autoencoder representations. We find a statistically significant enrichment of genome-wide association studies (GWAS) relevant genes in the last layer, and to a successively lesser degree in the middle and first layers respectively. In contrast, we find an opposite gradient where a modular proteinprotein interaction signal is strongest in the first layer, but then vanishing smoothly deeper in the network. We conclude that a data-driven discovery approach is sufficient to discover groups of disease-related genes. Code: gitlab.com/Gustafsson-lab. Full Paper: www.nature.com/articles/s41467-020-14666-6

1. Paper Summary

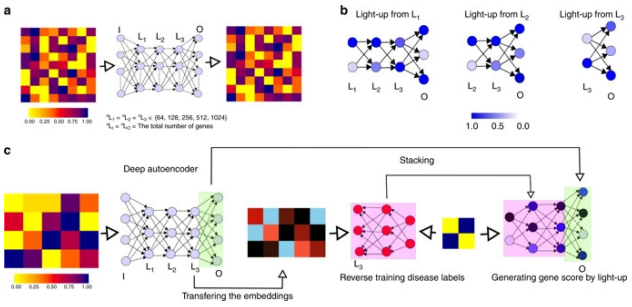

Figure 1. Schematic diagram of interpreting an autoencoder and defining the disease modules. a Training an autoencoder. b The steps of light-up method used for interpreting the hidden layer nodes in terms of PPI and pathways. c Depicts the steps of predicting the disease gene using transcriptomics signals and autoencoder.

References

Dwivedi, S. K., Tjärnberg, A., Tegnér, J., and Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nature Communications*, 11(856):1–10, 2020. https://doi.org/10.1038/s41467-020-14666-6.

[1]Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden [2]Department of Biology, Center For Genomics and Systems Biology, New York University, New York, NY 10008, USA [3]Center for Developmental Genetics, Department of Biology, New York University, New York, NY, USA [4]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955–6900, Saudi Arabia [5]Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden [6]Science for Life Laboratory, Solna, Sweden. Correspondence to: Mika Gustafsson <mika.gustafsson@liu.se>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).