うつ病とその症状の生物学的背景を解明するための機械学習アルゴリズムの開発

AUTHOR(S):

Petschner, Peter

CITATION:

Petschner, Peter. うつ病とその症状の生物学的背景を解明するための機械学習アルゴリズムの開発. 京都大学化学研究所スーパーコンピュータシステム研究成果報告書 2023, 2022: 16-16

ISSUE DATE:

2023-03

URL:

http://hdl.handle.net/2433/283138

令和4年度　京都大学化学研究所　スーパーコンピュータシステム　利用報告書

うつ病とその症状の生物学的背景を解明するための機械学習アルゴリズムの開発

The development of machine learning algorithms to decipher the biological background of major depression and its symptoms

Bioinformatics Center (Mamitsuka laboratory), Institute for Chemical Research, Kyoto University, Peter Petschner

研究成果概要

During FY2022/2023 we used a large, UK Biobank derived dataset to test a machine learning model and identify relevant genetic factors behind depression. The dataset contained 168,096 individuals, 2 environmental (phenotypic variables) factors, depression score as output and 3,792,532 genetic factors, in the form of so called single-nucleotide polymorphisms (SNPs).

Our original network architecture followed a feed-forward neural network with extensions to increase interpretability. Run of this model and its extension using side information about the ca. 3.5M SNPs needs the utilization of the SDF nodes of the supercomputer. After optimization steps data-loading and calculation times take about 7-9 minutes per 1000 individuals, memory use is 900GB-1.3TB. Thus, obtaining results still requires around 3 months of run time for a single run. We started up to December more than 17 runs, but many was killed by job manager. After we found solution with a system engineer jobs were running smoothly. The shutdown at the beginning of January caused small delays, due to modification of the code to allow for restarting of a job.

From a scientific perspective preliminary runs showed an average increase of 21.56%, if SNPs were added to sex and age, in the prediction performance of depression and aligned well with previous SNP-based heritability estimates of depression. The current interpretation method yielded 22 genes with consistent positive signal for depression. Among these, 3 (13.6%) showed overlap with the most recent genome-wide association study, giving 19 newly discovered genes.

The above indicate the potential of the new method, however, additional runs are necessary. Therefore, we continue the research and plan to extensively use the SC for the same purpose in the next fiscal year as well. Publications are expected in the future.