令和 5 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, Institute for Chemical Research, Kyoto University,
Peter Petschner

研究成果概要

During FY2023/2024 we continued the training, fine tuning and assessment of a neural network model on the UK Biobank derived dataset to identify genes behind major depressive disorder. The network architecture followed a feed-forward neural network of 3 layers with extensions to increase interpretability. As an extension to previous years, we began analysing the full-scale dataset containing 281,376 individuals, 2 environmental (phenotypic variables) factors, depression score as output and 3,778,895 genetic factors (single-nucleotide polymorphisms [SNPs]). We also used a smaller subset of this dataset, which consisted only 520,724 SNPs and the same number of individuals.

The two datasets were used to validate contributing genes. Nonetheless, while results from the small dataset showed consistent prediction performance improvement compared to a baseline, the model with the large dataset failed to achieve similar improvements. Initial assumptions were an increased amount of noise or unsatisfactory hyperparameter settings behind these results. For hyperparameter optimization and testing we inducted more the 20 runs with the large dataset using various settings, each requiring approximately 1.5 months of runtime and 900GB-1.3TB of memory, thus, these investigative steps consumed a large amount of time and resources. After several test runs we were still unable to achieve a performance improvement over the baseline running with the large dataset.

Following the contact and discussion with a consortium in the European Union working with UK Biobank data the validity of the large dataset became questionable. Careful investigation (independent to the supercomputer system) pointed to a potential inconsistency in the quality control steps of the data caused by a major bug in the most-widely used genetic software plink2 under default settings. The bug was eventually traced traced back and reported to the developers by our group (see version history entry on 5th of January, 2024 on https://www.cog-genomics.org/plink/2.0/). Restart of quality controls steps followed and re-runs are currently running for the remaining part of this fiscal year.

According to the above, publications are expected in the upcoming period once the final results are obtained.