令和 6 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

# 機械学習手法を使用した遺伝子のアノテーション付け
# Annotating genes using machine learning method

Kyoto University, Institute of Chemical Research, Ogata Lab　YU ZHAOXI

## 研究成果概要

Traditional sequence alignment methods like BLAST are too slow for annotating proteins in large datasets. To overcome this, I developed a machine learning model for KO annotation using the KEGG GENES database as training data. This model learns to assign annotations rapidly, eliminating the need for time-consuming sequence alignments.

Performance tests on CPU show that the machine learning approach is 34 times faster than BLAST and 6 times faster than the HMM-based tool KofamScan. With GPU acceleration, speed improvements can reach up to 1000 times, making it possible to annotate 40 million protein sequences from metagenomic samples in just 30 minutes.

These promising results highlight the potential of machine learning to transform protein annotation in high-throughput metagenomic studies. Future work will focus on further testing, optimization, and comprehensive comparisons with existing methods, with the ultimate goal of publishing these findings in a peer-reviewed journal.