

DeepKOALA: KEGG オーソロジー (KO) 割り当てのための高速かつ高精度な深層学習
フレームワーク

DeepKOALA: A Fast and Accurate Deep Learning Framework for KEGG Orthology
Assignment

京都大学 化学研究所 バイオインフォマティクスセンター 化学生命科学領域 余 兆熙

研究成果概要

Traditional sequence alignment approaches such as BLAST remain a major bottleneck for protein functional annotation at metagenomic scale. In response, we developed DeepKOALA, a deep learning framework that formulates KEGG Orthology (KO) assignment as an open-set recognition problem. Trained on the KEGG GENES database and implemented with a lightweight GRU backbone, DeepKOALA not only assigns KOs accurately for known functions but also rejects out-of-distribution sequences to reduce false annotations. To better match real metagenomic conditions, we further incorporated dedicated modules for fragmented sequences and multi-domain proteins, improving robustness in large and complex datasets.

In benchmarking experiments, DeepKOALA achieved an F1-score of 0.8653 for KO assignment, comparable to alignment-based annotation tools, while providing a substantial speed advantage. In particular, DeepKOALA delivered a $37.5\times$ acceleration over BlastKOALA, enabling scalable, high-throughput functional genomics workflows on shared computing infrastructure. The tool is now publicly available via GenomeNet (<https://www.genome.jp/tools/deepkoala/>) and the GitHub repository (<https://github.com/zhaoxi120/deepkoala>). A preprint has been released on bioRxiv (<https://www.biorxiv.org/content/10.64898/2026.01.07.698072v1-0>), and the manuscript is currently under submission to a peer-reviewed journal.