[BIOINFO 2025]

Clustering Proteomes and Curating References at Scale with MMseqs2-ProteomeCluster

Gyuri Kim[1], Martin Steinegger[1,2*]

[1] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea*
[2] *School of Biological Sciences, Seoul National University, Seoul, South* Korea.
**Corresponding author: martin.steinegger@snu.ac.kr*

The unprecedented growth in sequenced genomes submitted to public databases has driven a rapid increase in protein sequences and in new proteomes. In particular, the sequencing of nearly identical bacterial genomes as well as of (sub)strains with minimal genomic divergence, has caused great redundancy in UniProtKB. This redundancy primarily results in an inflated number of proteomes across all organisms, limiting the utility of these datasets. It also complicates the identification of reference proteomes, which traditionally relies on computationally intensive all-vs.-all pairwise alignments across proteomes.

Here we introduce MMseqs2-ProteomeCluster, which reduces redundancy by computing a reference proteome for any given species and identifying a subset of reference proteomes that cover most of the sequence space within that species. We first utilize MMseqs2-Linclust to rapidly cluster the proteins from all proteomes in linear time. The reference proteome is then constructed by including the most informative proteins shared among all proteomes and curated to maximize the coverage of the sequence space. Using this reference proteome, we compute pairwise alignments within clusters to group covered subsets and iteratively select additional references until every proteome of the species is fully represented.

We show this approach is orders of magnitude faster than standard workflows, processing 1.66M sequences from 288 proteomes of the genus Mycolicibacterium in 1 minute, compared to 16 hours with the method provided by UniProt. MMseqs2-ProteomeCluster is available as a module of MMseqs2 at https://mmseqs.com.