[BIOINFO 2023]

Petasearch: Efficient and Sensitive Sequence Comparison at Scale

Milot Mirdita[1], Minghang Li[1], Jonas Hügel[2], Johannes Söding[3,4*], Martin Steinegger[1,5,6,7*]

[1] *School of Biological Sciences, Seoul National University, Seoul, Korea*
[2] *Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Germany*
[3] *Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany*
[4] *Campus-Institut Data Science (CIDAS), Göttingen, Germany*
[5] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea*
[6] *Artificial Intelligence Institute, Seoul National University, Seoul, Korea*
[7] *Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea*
* *Corresponding authors: soeding@mpinat.mpg.de, martin.steinegger@snu.ac.kr*

The Sequence Read Archive is the central repository for genomics experiments and a treasure trove of over 70 petabases of sequence data. However, its massive size presents a significant challenge to traditional search methods. Time- and space-efficient search data structures, such as the Bloom-filter and sketching-based methods have been proposed as scalable alternatives, but their sensitivity is limited.

We present Petasearch, a tool for quickly and accurately searching protein sequences within large databases. Petasearch's algorithm involves three stages: First, the sequences in the database are pre-processed, sorted, and stored in a compressed k-mer index. Next, query k-mers are extracted and matched with the database k-mers, filtering out non-homologous sequences early. Finally, high-scoring k-mer matches are aligned with a SIMD-accelerated banded Smith-Waterman.

We optimize Petasearch using modern CPU caching and prefetching, advanced Linux IO techniques, and high read-bandwidth NVMe-SSDs. When tested across 21 NVMe-SSDs, Petasearch was found to be 15 and 145 times faster than current search algorithms for a 450GB and 9.3TB dataset, respectively. Petasearch maintains comparable sensitivity to state-of-the-art algorithms, detecting sequence identities as low as 60%, and identifying homology using its profile-search down to 40% in a SCOP25 benchmark.

Petasearch is available at petasearch.mmseqs.com as free open-source software for analysis and comparison of protein sequences at scale.