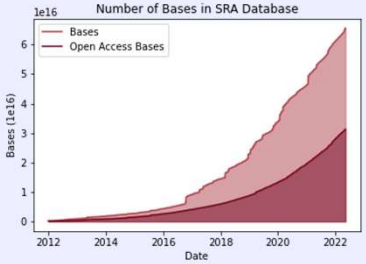
Petasearch: Fast, approximate comparison of huge sequence datasets

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

[*]Contributed equally, [1]School of Biological Sciences, Seoul National University, Seoul, South Korea, [2]Quantitative and Computational Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany, [3]Department of Medical Informatics, The University Medical Center Göttingen, Göttingen, Germany, [4]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, South Korea, [5]Artificial Intelligence Institute, Seoul National University, Seoul, South Korea,,

Abstract

The Sequence Read Archive (SRA) currently holds over 60 petabases and is growing rapidly in size.



This vast amount of data represents a treasure trove for medicine and biotechnology. Bloom-filter and sketching based approaches to find k-mer seeds were proposed to accelerate searches, however they offer only limited sensitivity.

We developed **petasearch** to enable fast and sensitive searching for proteins extracted from huge databases. Its algorithm contains three stages:

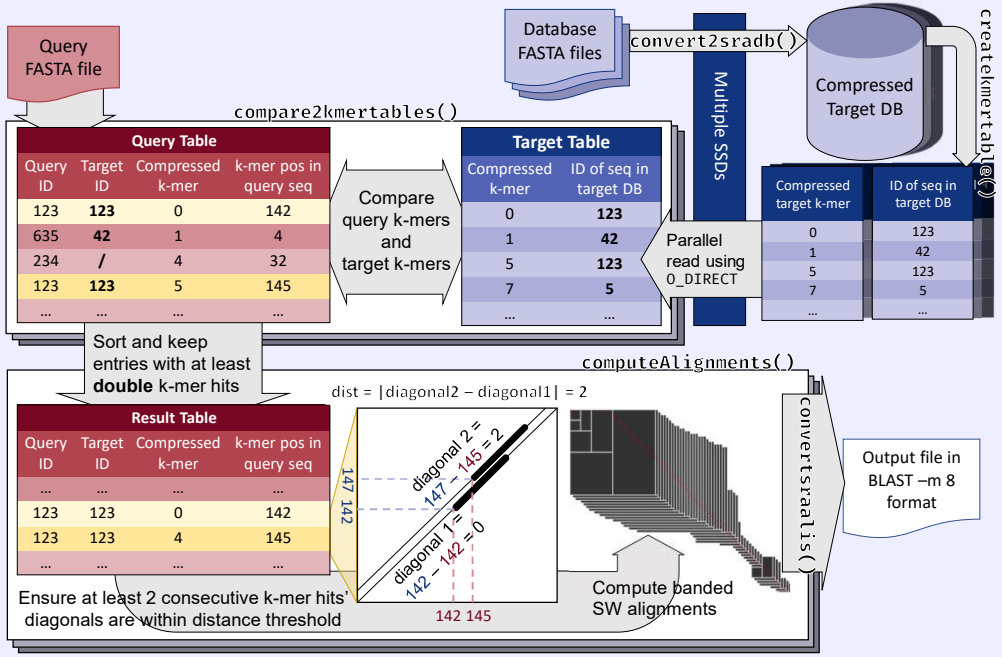
1. We pre-process the database sequences to extract k-mers, sort them and store them in a highly compressed k-mer index.
2. We extract query k-mers, add similar k-mers and find matches between query and database k-mers. We remove matches with only a single shared k-mer to reject non-homologous sequences early. To maximize throughput, we exploit the caching and prefetch infrastructure of modern CPUs, advanced Linux IO techniques, as well as the enormous read bandwidth of NVMe-SSDs.
3. We compute SIMD-accelerated banded Smith-Waterman alignments between sequences of high-scoring k-mer matches.

We show that **petasearch** is **190x** faster while querying a **9.3TB** dataset on **21** NVMe-SSDs. At much accelerated search speeds, **petasearch** matches state-of-the-art algorithms on sensitivity down to sequence identities of **60%**. On a SCOP25 benchmark we show that **petasearch**'s profile search detects sequence homology down to at least **40%** sequence identity.

Code Availability

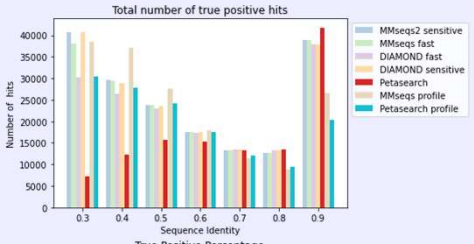
petasearch is a GPLv3-licensed open-source software available at https://github.com/steineggerlab/petasearch

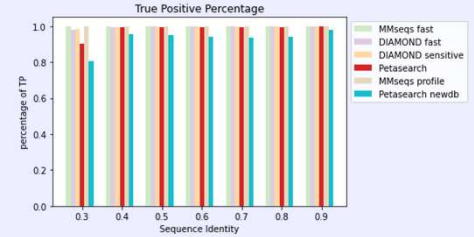
Architecture Overview



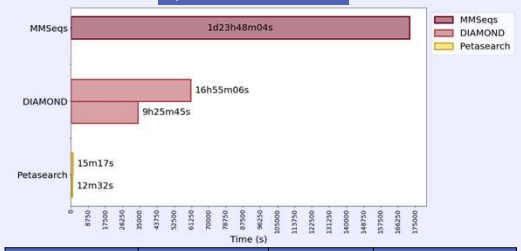
Sensitivity and Speed Benchmarks

Sensitivity Benchmark



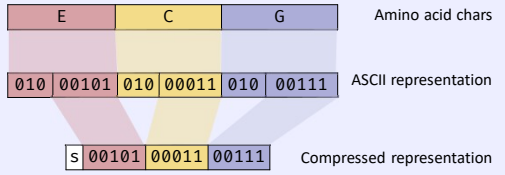


Speed Benchmark

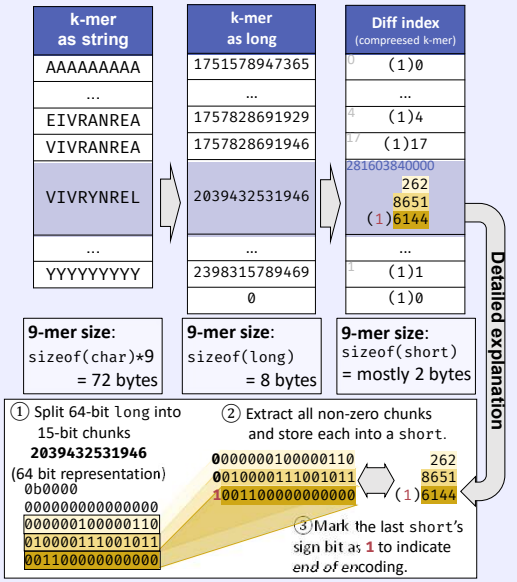


Name	Time	Speed up	Notes
petasearch	12m 32s	-	Exact k-mer
	15m 17s	-	Similar k-mer generation
DIAMOND	9h 25m 45s	**38x**	Only query-indexed
	16h 55m 06s	**68x**	Double-indexed query and target
MMseqs2	47h 48m 04s	**191x**	Default params

Sequence Compression



K-mer Index Compression



Conclusion

- **petasearch** is a lightning fast searching algorithm with comparable sensitivity at higher sequence identity.
- **petasearch** will enable researchers to exploit the vast amount of evolutionary information available in current and upcoming databases.

Acknowledgement

This work is supported by XXX research fund program. XXXXXXXXX