Quantum leap in fast and deep protein sequence similarity searching

Martin Steinegger and Johannes Söding

Sequence similarity searching is widely used in life sciences to infer the functions and structures of query proteins from similar proteins in sequence databases. In the booming field of metagenomics, huge amounts of environmental sequences need to be annotated, but often search tools find no matches. To address this gap, Martin Steinegger and Johannes Söding at the MPI-BPC have developed an open-source software for fast similarity searching and clustering of protein sequences. In its iterative profile search mode, MMseqs2 (Many-against-Many sequence searching) drastically improves the sensitivity and the speed over current state-of-the-art methods.

Metagenomics revolution

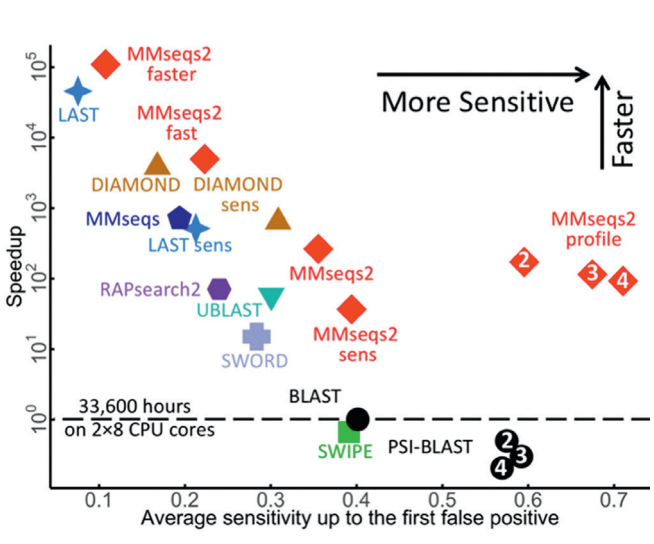
Metagenomics is revolutionizing the study of microbes in their natural environments, such as the human gut, the oceans, or soil, and is revealing the enormous impact of microbes on our health, our climate, and ecology. In metagenomics, DNA or RNA of bacteria, archaea, and viruses are sequenced directly, making the 99 % of microbes amenable to investigation that cannot be cultivated in the lab. Combined with the enormous drop in sequencing costs by a factor of ten thousand in just ten years, this has led to an explosive growth in the amount of sequence data in public databases.

To predict functions for these new sequences, very fast sequence search tools have been developed in recent years, but the increase in speed was paid by lower search sensitivity. Yet many of the microbes investigated by metagenomics have no close relatives in the sequence databases, and current search tools are too insensitive to detect them. Consequently, for the large majority of metagenomic sequences no functions can be predicted.

More sensitive than PSI-BLAST and 400 times faster

To address the need for very fast yet sensitive sequence search, Steinegger and Söding developed the software MMseqs2. The most important distinction of MMseqs2 to previous fast search tools is its ability to search with sequence profiles and not only with simple sequences. Since PSI-BLAST made its debut 20 years ago, sequence profiles have been known to improve search sensitivity enormously. But until now, no way had been found to drastically speed up sequence profile searches.

This changed with the new, very fast, and sensitive sequence prefilter algorithm at the core of MMseqs2. It preselects the most promising database sequences for subsequent, slower, and more accurate comparison. Whereas all recent tools use *exact* matches between short words



Average search sensitivity versus relative speed for various fast sequence search tools. White numbers inside the plot symbols give the number of iterations of sequence profile searches with MMseqs2 and PSI-BLAST.

(*k*-mers) of amino acid letters, the researchers extend a 27 year-old idea from BLAST to detect *similar* instead of exact *k*-mer matches. Crucially, their algorithm can generate lists of similar *k*-mers both for sequences and sequence profiles. To gain further sensitivity, they were able to increase the word length *k* from three to seven and also developed a very time-efficient method to detect when two neighboring *k*-mer matches occur at just the right spacing that a similarity by pure chance is unlikely.

MMseqs2 scales almost inversely with the number of used processor cores. It can automatically split and distribute query or target databases across several servers, allowing even users with relatively modest computing resources to cluster or search databases with billions of sequences. It also enables users to analyze jointly collections of datasets that could so far only be analyzed separately.

"I am confident that because MMseqs2 addresses the pressing need for higher search speed and sensitivity it will become the standard tool for fast protein sequence searching", Söding concludes.

Original publication

Steinegger M, Söding J: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).