Reducing the computational power required to analyze DNA

July 29 2021

Credit: CC0 Public Domain

An approach that reduces the computational power required to analyze huge amounts of DNA data for identifying new microbes and their proteins could be used for manufacturing anything from new antibiotics to plastic-degrading enzymes.

An online platform now gives scientists around the world access to KAUST's advanced computational resources to improve understanding of the kinds of microbes that exist in different environments, and what they can do. The tool is expected to help researchers identify proteins and enzymes that can be used in agriculture, pharmaceuticals, the energy sector and many other industries.

Preparing bacterial cultures is routine: a scientist takes a sample from a wound, for example, and grows bacteria from it in laboratory petri dishes. The problem is that 99 percent of bacteria in these and other samples cannot be cultured like this in the laboratory. This makes it extremely difficult to discover the estimated one trillion microbes that exist.

To overcome this problem, scientists introduced an approach in 1998 called metagenomics sequencing, where a sample, such as a bucket of seawater, is taken from any environment and then analyzed for DNA. Scientists apply a method called shotgun sequencing that fragments any DNA in the sample into smaller pieces called reads. These metagenomic short reads are then reassembled to identify genes. Over the years, a tremendous amount of microbial sequencing data has been extracted from different environments, but analysis requires state-of-the-art methods, recent reference databases and huge computational prowess.

To address this problem, Intikhab Alam, Vladimir Bajic, Carlos Duarte, Takashi Gojobori and colleagues, developed the KAUST Metagenomic Analysis Platform (KMAP). "Using KMAP, we were able to analyze and compare 275 million microbial genes in only 13 days using KAUST's Shaheen II supercomputer. In comparison, this would have required 522 years using a single computer CPU," says Alam.

The process involves first assembling short sequencing reads into longer contigs or assemblies using state-of-the-art metagenomics assembly

tools. It is this data that can be input into KMAP's annotation module. Scientists can either input their own assembled contigs, genes or gene catalogs into the platform or analyze and compare existing KMAP-annotated data from several habitats. For easy and interactive analyses, KMAP-annotated gene information tables can be sifted through using KMAP's compare module to gain deeper insight into the types of microbes found in different environments and their functions.

The KAUST team used the data they assembled to find microbial enzymes that could be used to degrade plastic waste in the oceans. They also sifted through the data to identify antibiotic-resistance genes in bacteria that live in soil and underwater thermal vents.

"KMAP will give researchers across the world equal access to data processed through KAUST's advanced computational resources and eliminate the need for advanced bioinformatics skills in order to explore microbial communities and functions," says Gojobori.

   More information: Intikhab Alam et al, KAUST Metagenomic Analysis Platform (KMAP), enabling access to massive analytics of re-annotated metagenomic data, *Scientific Reports* (2021). DOI: 10.1038/s41598-021-90799-y