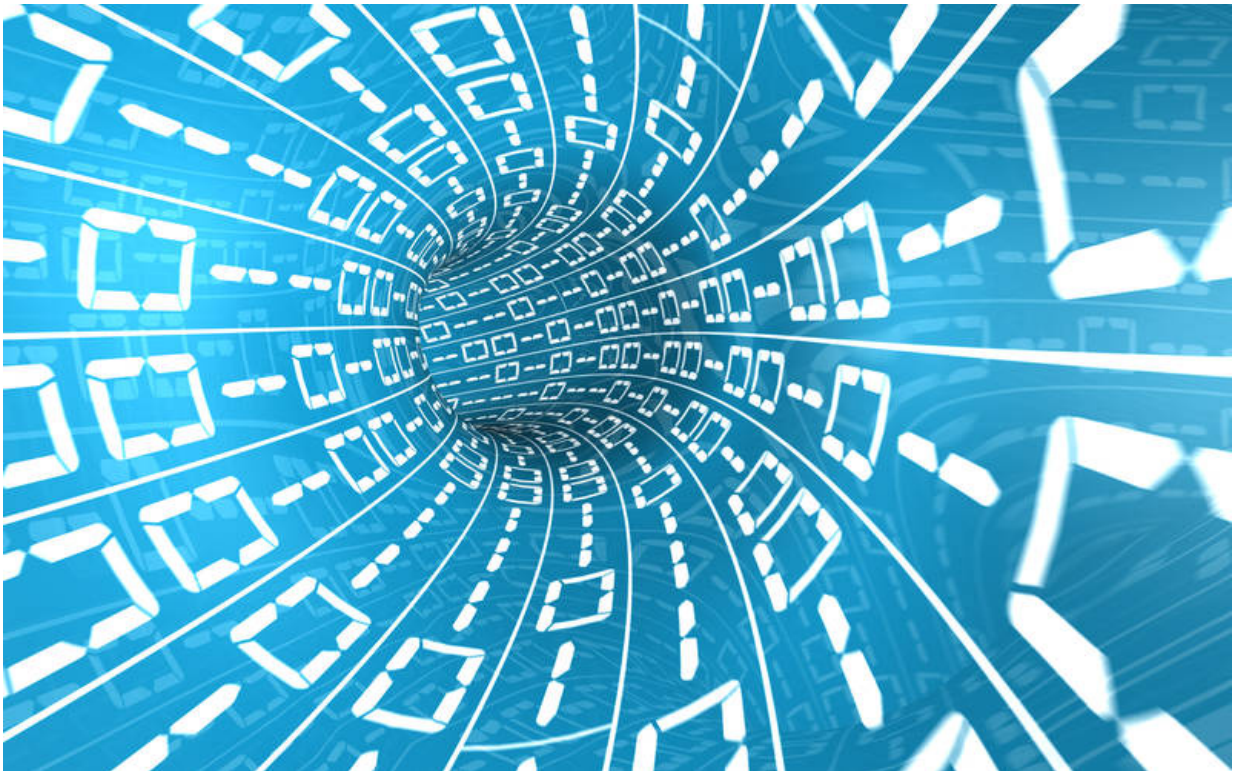# Big data processing enables worldwide bacterial analysis

October 6 2016

The Sequence Read Archive, a public database for deposition of sequences, currently stores over 100,000 gene sequence datasets which previously could not be evaluated in their whole. Credit: Fotolia/ Dreaming Andy

Sequencing data from biological samples such as the skin, intestinal tissues, or soil and water are usually archived in public databases. This

allows researchers from all over the globe to access them. However, this has led to the creation of extremely large quantities of data. To be able to explore all these data, new evaluation methods are necessary. Scientists at the Technical University of Munich (TUM) have developed a bioinformatics tool which allows to search all bacterial sequences in databases in just a few mouse clicks and find similarities or check whether a particular sequence exists.

Microbial communities are essential components of ecosystems around the world. They play a key role in key biological functions, ranging from carbon to nitrogen cycles in the environment to the regulation of immune and metabolic processes in animals and humans. That is why many scientists are currently investigatin microbial communities in great detail.

Sequencing for microbiological DNA analysis

The Sanger sequencing method developed in 1975 used to be the gold standard to decipher the DNA code for 30 years. Recently, next generation sequencing technologies, or NGS as they are called, have led to a new revolution: With minimal personnel requirements, current devices can, within 24 hours, generate as much data as a hundred runs of the very first DNA sequencing method.

Today, the sequencing analysis of bacterial 16S rRNA genes is the most frequently used identification method for bacteria. The 16S rRNA genes are seen as ideal molecular markers for reconstructing the degree of relationship between organisms, as their sequence of nucleotides (the building blocks of DNA) has been relatively conserved throughout evolution and can be used to infer phylogenetic relationships between microorganisms. The acronym rRNA stands for ribosomal ribonucleic acid.

The Sequence Read Archive (SRA), a public database for deposition of sequences, currently stores over 100,000 such 16S rRNA gene sequence datasets. This is because the new technical procedures for DNA sequencing have caused the volume and complexity of genome research data over the past few years to grow exponentially. The SRA is home to datasets which previously could not be evaluated in their whole.

"Over all these years, a tremendous amount of sequences from human environments such as the intestine or skin, but also from soils or the ocean has been accumulated", explains Dr. Thomas Clavel from the Institute for Food and Health (ZIEL) at the TU Munich. "We have now created a tool which allows these databases to be searched in a relatively short amount of time in order to study the diversity and habitats of bacteria", says Clavel—"with this tool, a scientists can conduct a query within a few hours in order to find out in which type of samples the bacterium he is interested in can be found—for example a pathogen from a hospital. This was not possible before." The new platform is called Integrated Microbial Next Generation Sequencing (IMNGS) and can be accessed via the main website http://www.imngs.org.

A detailed description of how IMGS functions using the intestinal bacterium Acetatifactor muris has been published in the current online issue of "Scientific Reports". Registered users can carry out queries filtered by the origin of the bacterial data, or also download entire sequences.

Such bioinformatics approach may soon become indispensable in routine daily clinical diagnostics. However, one critical aspect is that many members of complex microbial communities remain to be described. "Improving the quality of sequence datasets by collecting new reference sequences is a great challenge ahead", says Clavel—"moreover, the quality of datasets is not yet good enough: the description of individual samples in databases is incomplete, and hence the comparison

possibilities using IMNGS are currently still limited."

However, Clavel imagines that a collaboration with clinics could be a catalyst for progress, provided the database is filled more meticulously. "If we had very well-maintained databases, we could use innovative tools such as IMNGS to possibly help diagnosis of chronic illnesses more rapidly", says Clavel.

Provided by Technical University Munich

Citation: Big data processing enables worldwide bacterial analysis (2016, October 6) retrieved 17 July 2024 from https://phys.org/news/2016-10-big-enables-worldwide-bacterial-analysis.html