In the last GovMath we highlighted two examples of interdisciplinary teams coming together to solve a complex problem. We continue that theme now with bioinformatics and molecular biology researchers at Lawrence Livermore National Laboratory developing algorithms and computational strategies to analyze microbial communities previously unable to be explored due to the high complexity in the samples. The discoveries enabled by their advances reach as far as the International Space Station!

Profiling Microbial Communities with Shotgun Metagenomics

Jonathan E. Allen and Crystal Jaing

The diversity of microbial life remained largely unexplored until recent advances in DNA sequencing technology. The advent of shotgun metagenomics enables the capture of microbial communities that were previously inaccessible through traditional cultured-based methods that fail to grow most microbes in a laboratory environment or low-content molecular assays that can only interrogate one organism or gene at a time. A key innovation of shotgun metagenomics is the ability to examine a biological sample without relying on previously identified genetic sequences, which gives a much less biased reporting on the true genetic content of a microbial community. This new and more complete view of microbial diversity through understanding the genomics is giving a better understanding of the importance of the microbiome on human health and the environment. Shotgun metagenomics presents the challenge of reconstructing genomes via the collection of fragments of genomes. Since we start out with a community rather than a homogenous sample, it is difficult to accurately assign membership for each genetic fragment to an organism, particularly when the organism may not previously have been seen. The challenge is further compounded by the need to collect tens of millions of genetic fragments in order to adequately sample the microbial community.

One strategy that has had success in recent years is to apply a comprehensive "k-mer" based search between the newly recovered genomic fragments and all previously sequenced organisms. Our own novel k-mer based approach is demonstrated through the Livermore Metagenomic Analysis Toolkit (LMAT), an open source and highly scalable metagenomic analysis tool (Ames et al., 2015). Each genetic fragment is represented as a string of nucleotides and decomposed into overlapping k-mers, substrings of fixed length k. (Typically k is set to a value between 20 and 32.) The k-mers of previously sequenced genomes are stored in a searchable index to track the organism membership for each k-mer. A hierarchical taxonomic tree is used to track k-mers that belong to multiple organisms. Each genetic fragment is searched against the index to assign membership to either a single organism or a family of organisms

Jonathan Allen is an informatics team lead in the Global Security Applications Division of the Computation Directorate at Lawrence Livermore National Laboratory. His email address is allen99@llnl.gov.

Crystal Jaing is a molecular biologist and the applied genomics group leader in the Biosciences and Biotechnology Division of the Physical Life Sciences Directorate at Lawrence Livermore National Laboratory. Her email address is jaing2@llnl.gov.

Communicated by Notices *Associate Editor Emilie Purvine.*

DOI: https://dx.doi.org/10.1090/noti1960

depending on genetic uniqueness. Some k-mers will match to too many genomes, which increases the computational cost of selecting an organism from two closely related organisms. A balance must be struck between retaining all genome information to make individual organism calls and increasing search speed by not differentiating closely related species or subspecies. The algorithm applies a tunable pruning procedure, which removes taxonomic levels in the tree for each k-mer to balance the need for precise individual genome calls with search speed. For example, when a k-mer matches to a thousand individual genomes, the genome information would be pruned from the tree and moved up to the species level. If the k-mer continues to map to too many species the pruning continues until the k-mer can be assigned to a small number of taxonomic variants such as genus or family. Analysis from LMAT allows genetic sequences from complex samples to balance accurately mapping to existing genomes of organisms with the need for rapid analysis, thus providing a new level of accuracy to the profile of the composition of that community and possible insights into its functional properties.

A key feature of LMAT is the large database of microbial and eukaryotic sequences, enabling sensitive and specific mapping of genetic contents from very complex metagenomic data sets. While the search algorithm is optimized for speed by limiting redundant examination of k-mers that occur in many reference genomes, there is the added cost to store an index of hundreds of gigabytes in size, which can exceed the capacity of traditional fast computer memory (Dynamic Random Access Memory or DRAM). We significantly reduced the cost by optimizing the index to use newer disk storage (Non-Volatile Random Access Memory or NVRAM) as a supplemental memory resource. The search algorithm limits the number of times genome information must be retrieved from a slower disk by storing the more frequently accessed portions in fast memory. This allows the search of a much larger collection of previously sequenced genomes in parallel, which would not be possible by relying on traditional memory alone. These methods have been used to profile microbes in a wide array of settings and provide new insights on microbial communities, including detecting potential etiologic disease agents (Thissen et al., 2018) and microbial communities on the International Space Station (Be et al., 2017). The results show the need to account for microbial diversity when studying contributing factors to human health.

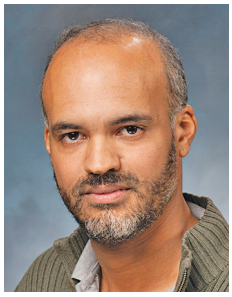
This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

References

Ames SK, Gardner SN, Marti JM, Slezak TR, Gokhale MB, Allen JE. Using populations of human and microbial genomes for organism detection in metagenomes, *Genome Res.* **25** (2015), 1056–1067.

Be NA, Avila-Herrera A, Allen JE, Singh N, Checinska Sielaff A, Jaing C, Venkateswaran K. Whole metagenome profiles of particulates collected from the International Space Station, *Microbiome* **5** (2017), 81.

Thissen JB, Isshiki M, Jaing C, Nagao Y, Aldea DL, Allen JE, Izui M, Slezak TR, Ishida T, Sano T. A novel variant of torque teno virus 7 identified in patients with Kawasaki disease, *PLOS ONE* **13** (2018), e0209683.



Jonathan E. Allen

Crystal Jaing

Credits

Author photos are courtesy of Lawrence Livermore National Laboratory.