Engineers develop innovative microbiome analysis software tools

May 7 2024

Credit: CC0 Public Domain

Since the first microbial genome was sequenced in 1995, scientists have reconstructed the genomic makeup of hundreds of thousands of microorganisms and have even devised methods to take a census of

bacterial communities on the skin, in the gut, or in soil, water and elsewhere based on bulk samples, leading to the emergence of a relatively new field of study known as metagenomics.

Parsing through metagenomic data can be a daunting task, much like trying to assemble several massive jigsaw puzzles with all of the pieces jumbled together. Taking on this unique computational challenge, Rice University graph-artificial intelligence (AI) expert Santiago Segarra and computational biologist Todd Treangen paired up to explore how AI-powered data analysis could help craft new tools to supercharge metagenomics research.

The scientist duo zeroed in on two types of data that make metagenomic analysis particularly challenging—repeats and structural variants—and developed tools for handling these data types that outperform current methods.

Repeats are identical DNA sequences occurring repeatedly both throughout the genome of single organisms and across multiple genomes in a community of organisms.

"The DNA in a metagenomic sample from multiple organisms can be represented as a graph," said Segarra, assistant professor of electrical and computer engineering.

"Essentially, one of the tools we developed leverages the structure of this graph in order to determine which pieces of DNA appear repeatedly either across microbes or within the same microorganism."

Dubbed GraSSRep, the method combines self-supervised learning, a machine learning process where an AI model trains itself to distinguish between hidden and available input, and graph neural networks, systems that process data representing objects and their interconnections as

graphs.

The paper, also available on the *arXiv* preprint server, was presented at the 28th session of an annual international conference on research in computational molecular biology, RECOMB 2024. The project was led by Rice graduate student and research assistant Ali Azizpour. Advait Balaji, a Rice doctoral alumnus, is also an author on the study.

Repeats are of interest because they play a significant role in biological processes such as bacterial response to changes in their environment or microbiomes' interaction with host organisms. A specific example of a phenomenon where repeats can play a role is antibiotic resistance.

Generally speaking, tracking repeats' history or dynamics in a bacterial genome can shed light on microorganisms' strategies for adaptation or evolution. What's more, repeats can sometimes actually be viruses in disguise, or bacteriophages. From the Greek word for "devour," phages are sometimes used to kill bacteria.

"These phages actually show up looking like repeats, so you can track bacteria-phage dynamics based off the repeats contained in the genomes," said Treangen, associate professor of computer science.

"This could provide clues on how to get rid of hard-to-kill bacteria, or paint a clearer picture of how these viruses are interacting with a bacterial community."

Previously when a graph-based approach was used to carry out repeat detection, researchers used predefined specifications for what to look for in the graph data. What sets GraSSRep apart from these prior approaches is the lack of any such predefined parameters or references informing how the data is processed.

"Our method learns how to better use the graph structure in order to detect repeats as opposed to relying on initial input," Segarra said. "Self-supervised learning allows this tool to train itself in the absence of any ground truth establishing what is a repeat and what is not a repeat. When you're handling a metagenomic sample, you don't need to know anything about what's in there to analyze it."

The same is true in the case of another metagenomic analysis method co-developed by Segarra and Treangen—reference-free structural variant detection in microbiomes via long-read coassembly graphs, or rhea. Their paper on rhea will be presented at the International Society for Computational Biology's annual conference, which will take place July 12–16 in Montreal.

The lead author on the paper is Rice computer science doctoral alumna Kristen Curry, who will be joining the lab of Rayan Chikhi—also a co-author on the paper—at the Institut Pasteur in Paris as a postdoctoral scientist. A version of the paper is available on the *bioRxiv* preprint server.

While GraSSRep is designed to deal with repeats, rhea handles structural variants, which are genomic alterations of 10 base pairs or more that are relevant to medicine and molecular biology due to their role in various diseases, gene expression regulation, evolutionary dynamics and promoting genetic diversity within populations and among species.

"Identifying structural variants in isolated genomes is relatively straightforward, but it's harder to do so in metagenomes where there's no clear reference genome to help categorize the data," Treangen said.

Currently one of the widely used methods for processing metagenomic data is through metagenome-assembled genomes or MAGs.

"These de novo or reference-guided assemblers are pretty well-established tools that entail a whole operational pipeline with repeat detection or structural variants' identification being just some of their functionalities," Segarra said.

"One thing that we're looking into is replacing existing algorithms with ours and seeing how that can improve the performance of these very widely used metagenomic assemblers."

Rhea does not need reference genomes or MAGs to detect structural variants, and it outperformed methods relying on such prespecified parameters when tested against two mock metagenomes.

"This was particularly noticeable because we got a much more granular read of the data than we did using reference genomes," Segarra said.

"The other thing that we're currently looking into is applying the tool to real-world datasets and seeing how the results relate back to biological processes and what insights this might give us."

Treangen said GraSSRep and rhea combined—building on previous contributions in the area—have the potential "to unlock the underlying rules of life governing microbial evolution."

The projects are the result of a yearslong collaboration between the Segarra and Treangen labs.

"This has been a product of performing multiyear collaborative research across different areas of expertise, which has allowed our students Ali and Kristen to challenge existing paradigms and develop new approaches to existing problems in metagenomics," Treangen said.

More information: Ali Azizpour et al, GraSSRep: Graph-Based Self-Supervised Learning for Repeat Detection in Metagenomic Assembly, *arXiv* (2024). DOI: 10.48550/arxiv.2402.09381

Kristen D. Curry et al, Reference-free Structural Variant Detection in Microbiomes via Long-read Coassembly Graphs, *bioRxiv* (2024). DOI: 10.1101/2024.01.25.577285

Provided by Rice University