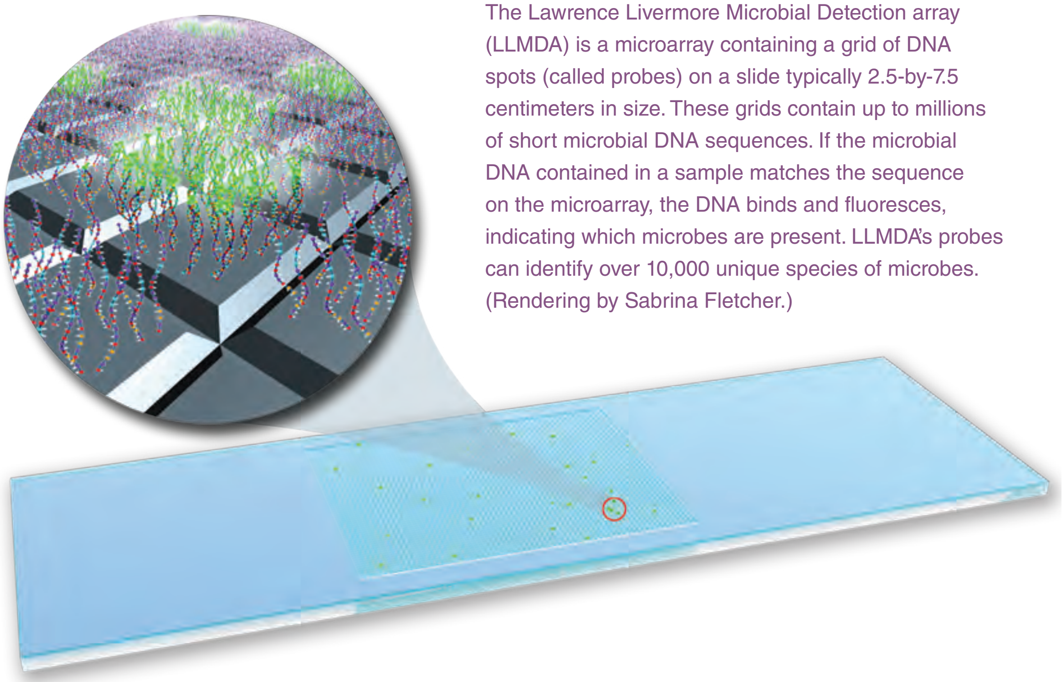
Two-Part Microbial Detection Enhances Bioidentification



The Lawrence Livermore Microbial Detection array (LLMDA) is a microarray containing a grid of DNA spots (called probes) on a slide typically 2.5-by-7.5 centimeters in size. These grids contain up to millions of short microbial DNA sequences. If the microbial DNA contained in a sample matches the sequence on the microarray, the DNA binds and fluoresces, indicating which microbes are present. LLMDA's probes can identify over 10,000 unique species of microbes. (Rendering by Sabrina Fletcher.)

A military medical team responding to an infantry soldier with extensive wounds must quickly deduce the most effective way to treat the victim. Typically, clinicians determine a wound-care approach based upon general intuition and experience. This approach must include considerations such as the optimal time to close the wound, how healing will be affected if the wound is closed immediately, which antibiotics to administer, and how other factors—including what bacteria may be in the wound—could help or hinder the patient's overall recovery.

New microbial detection technologies may help clinicians make more informed treatment decisions specific to a patient's needs. Laboratory researchers have recently demonstrated an approach that uses the Lawrence Livermore Microbial Detection Array (LLMDA) and the Livermore Metagenomic Analysis Toolkit (LMAT) for identifying and quantifying microbes to improve patient care. LLMDA offers a quick, cost-effective initial screening of a sample from a patient's wound. LMAT, on the other hand, provides more thorough examination of microbes by analyzing the exact order of chemical bases that form their DNA, called sequences, to identify pathogens and their genetic attributes.

"A perfect world of pathogen surveillance and diagnostics is analogous to a wheel and a spoke," explains LMAT principal investigator Jonathan Allen. "LLMDA is the wheel's hub—a high throughput screening tool that processes and categorizes the majority of information. The spokes are LMAT's sequencing methods, which provide deeper analysis of a sample, for identifying any initially undetected data or for gathering additional details." Used together, these two technologies are serving as the basis to transform biodefense efforts and improve medical treatments.

One Way or Another

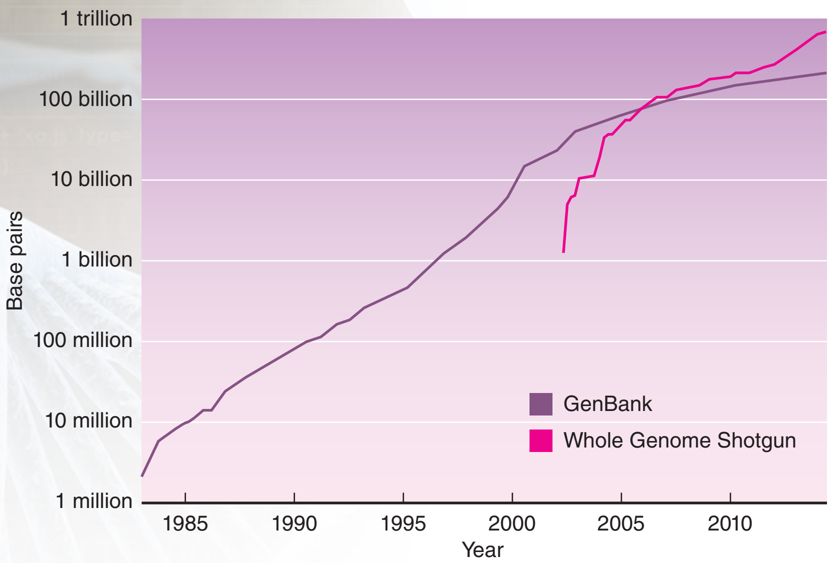
LLMDA is a microarray containing a grid of DNA spots on a 2.5-by-7.5-centimeter slide. (See *S&TR* April/May 2013, pp. 4–11). These grids contain hundreds of thousands to millions of short microbial DNA sequences, which are the complements of genetic markers found in microbes. If the microbial DNA contained in a fluid genetic sample matches the sequence on the microarray, the DNA binds and fluoresces, indicating which microbes are present. Traditional pathogen identification methods require up to several days to analyze a sample, but LLMDA boasts a 24-hour processing time. LLMDA can identify more than 10,000 unique species of microbes. However, the system can only detect previously sequenced microbes whose DNA complement is included on the array.

In contrast, LMAT delivers specific genetic information for microbes and identifies any that have been previously sequenced. LMAT's open-source software searches through genetic data catalogued in public repositories to more specifically identify organisms. In fact, a major accomplishment of LMAT is its ability to compile genetic variation into a searchable, scalable, and comprehensive database—a feat never before accomplished.

The vast amount of genetic information contained in microbial whole-genome repositories, currently totaling 115 gigabases and growing, posed significant challenges for the LMAT team. "Many bioinformatics problems are memory constrained because too much data is available," says computer scientist Maya Gokhale, who researches solutions to big-data problems in many fields, including metagenomics. These challenges will continue to grow as gene sequencing becomes easier, cheaper, and faster to execute, and as available genomic data increases exponentially. To overcome these memory and data storage problems, the LMAT team turned to innovations in supercomputing hardware and software.

Big Data, Big Memory

Gokhale and her team addressed LMAT's big-data issue by applying cutting-edge storage and access methods. "Our development approach to LMAT's metagenomic analysis and classification was centered upon anticipating the very large



As genome-sequencing processes become easier to execute, the amount of genome sequence data steadily increases with time. This graph shows sequenced genome data measured in base pairs over time, from two genome databases. The data influx poses a need for scalable software such as the Livermore Metagenomic Analysis Toolkit (LMAT), which enables users to identify pathogens including diseases, bacteria, viruses, fungi, and other organisms.

memory that would be required," says Gokhale. The flash drive, a type of nonvolatile random-access memory (NVRAM), provides a supplemental memory resource for retaining data even when the processor's power is off. Spreading database storage across NVRAM uses less main memory, called dynamic random access memory (DRAM). Unlike conventional storage structures that use disk memory, LMAT's use of both flash and main memories increases processing speeds.

LMAT was run on Livermore's Catalyst cluster to test the data storage and analysis approach. Catalyst is a first-of-its-kind architecture whose design was influenced by Gokhale and her team's memory storage research. The system is equipped with an impressive amount of both main and flash memories, which makes it ideally suited for solving big-data problems. LMAT was copied into the flash memory of each of Catalyst's 324 nodes. Individual nodes contain 800 gigabytes of flash memory alone.

However, the team's use of both main and flash memory made it difficult to retrieve data. "Traditionally, to access data, it has to be stored in DRAM," explains Gokhale. To retrieve needed data from both memories, Gokhale's team created a clever caching algorithm that transparently fetches a DNA sequence from NVRAM and moves it into DRAM, and from DRAM into the central processing unit.

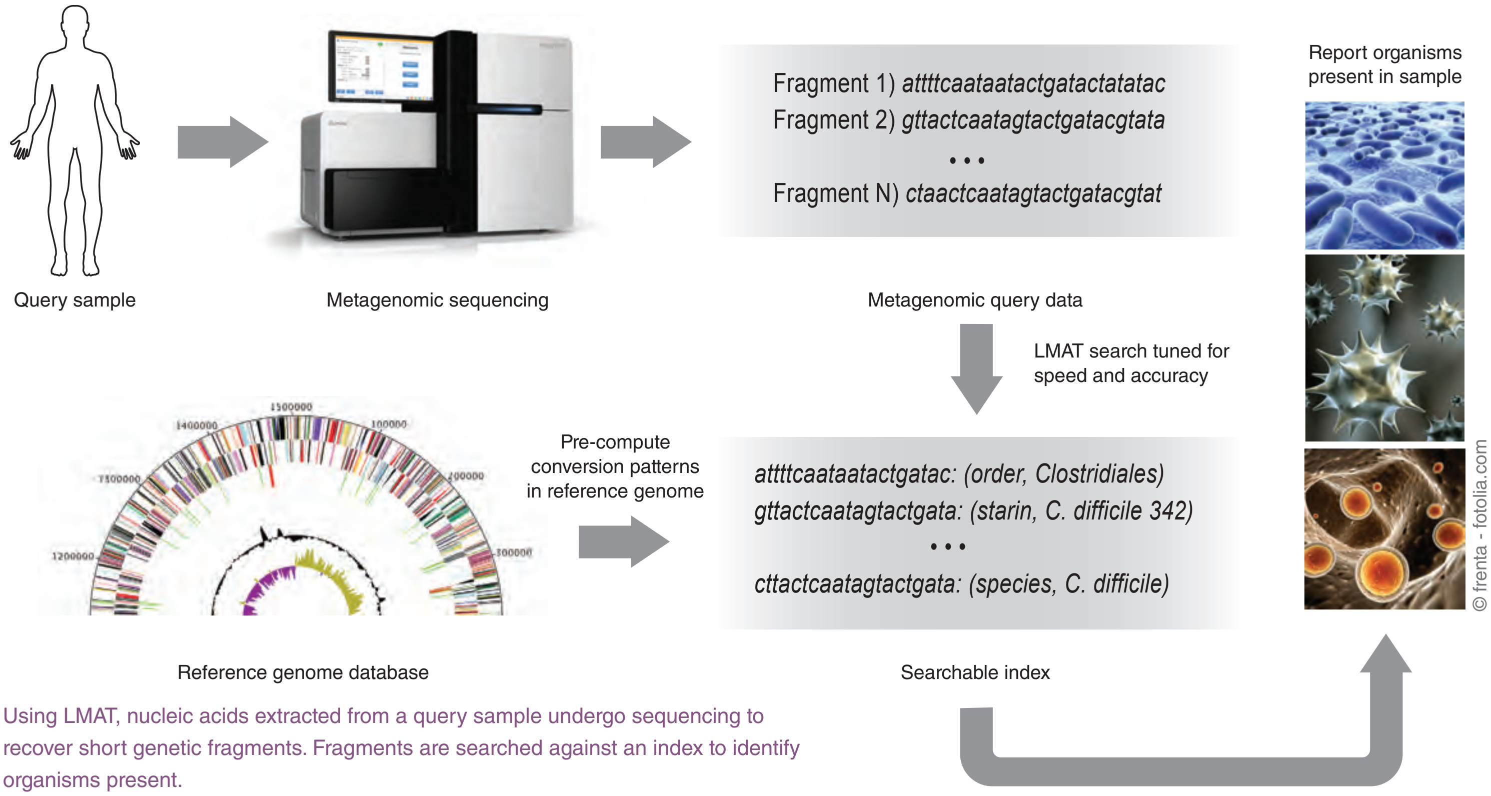
The caching system analyzes small sequences of genomic data called k-mers—continuous genetic sequences of length *k*. The team compiled all previously sequenced microbial genomes and broke them into short, contiguous sequences, where the k-mer length was 20 base pairs. Each k-mer was tagged by its source genome and marked according to its taxonomic level, which defines groups of biological organisms on the basis of shared characteristics. When LMAT identifies a k-mer, the database also indicates possible related organisms according to its taxonomic label. Looking at all the matching k-mers for a query sequence, Allen created an algorithm to select the taxonomic classification, which best reflects the taxonomic classification of the individual matching k-mers.

Furthermore, the team developed a method for faster retrieval of frequently fetched items. Index data is split into two components: data held in main memory and data held in flash memory. As an example, the first 10 base pairs of the k-mer may be cached in main memory if they have been searched for previously. K-mers similar to the sample in main memory then point to additional k-mers stored in flash memory that share the sample's first 10 base pairs. This complete process reduces trips to the flash drive, improving data retrieval time.

Catalyst has demonstrated LMAT's capabilities and is a proven asset for the approach. However, computers of its caliber are uncommon to other laboratories and universities, which limits LMAT's potential user base. "Our strength is our weakness," explains Allen. "Our index requires computing resources with a certain sophistication that is not readily accessible to many users." Thus, Allen is working to build an Amazon "instance," a virtual server containing LMAT, on Amazon's Elastic Compute Cloud for running applications via Amazon Web Services. Amazon account holders could download the instance and run LMAT remotely. "Our ultimate goal is to increase access to the technology," says Allen. "Those who do not have Livermore's computing resources or expertise should still be able to leverage LMAT's sequencing analysis."

The Treatment Two-Step

When used in conjunction, LLMDA and LMAT can help clinicians make more informed treatment decisions by identifying


Query sample
Metagenomic sequencing
Fragment 1) attttcaataatactgatactatatac
Fragment 2) gttactcaatagtactgatacgtata
• • •
Fragment N) ctaactcaatagtactgatacgtat
Metagenomic query data
Report organisms present in sample
LMAT search tuned for speed and accuracy
Pre-compute conversion patterns in reference genome
attttcaataatactgatac: (order, Clostridiales)
gttactcaatagtactgata: (starin, C. difficile 342)
• • •
cttactcaatagtactgata: (species, C. difficile)
Reference genome database
Searchable index
1500000
1400000
100000
1300000
200000
1200000
300000

© frenta - fotolia.com

Using LMAT, nucleic acids extracted from a query sample undergo sequencing to recover short genetic fragments. Fragments are searched against an index to identify organisms present.

specific bacteria in a wound that could affect the healing process. Laboratory researchers tested the dual approach on wound samples from 44 patients. LLMDA was used to assess 124 samples from combat wounds that had different healing outcomes. A subset of samples was then sequenced and processed by LMAT. In several cases, LMAT detected organisms LLMDA had not. The system could also potentially identify microbes that harbor antibiotic-resistant genes—a testament to its sensitivity of sequencing and LMAT's proficiency for detecting organisms in minute amounts.

"The combination of both technologies offers an effective workflow," explains microbiologist Nicholas Be. The team identified all present microbes—bacteria, viruses, and fungi—in the healed versus unhealed samples to compare microbial influence on the overall outcome. Results indicated that the presence of certain bacteria in the wound was associated with either positive or failed healing. Some types of bacteria, such as *Escherichia coli* and *Salmonella*, commonly found in the human gastrointestinal tract, were more frequently observed in wounds that healed successfully. This result defies traditional assumptions that all bacteria are malignant and should be eliminated and supports the notion that more specific microbial information better guides wound-treatment decisions.

"Thanks to these Livermore-developed technologies," says Be, "our team determined that it's insufficient to merely look at a wound for microbe presence or wound cleanliness and base treatment on that assessment." These findings could lead to improved wound-treatment processes in the clinic as well as in the field, as metagenomics researchers are working to make microbial detection technologies portable.

LLMDA and LMAT are products of an impressive combination of unique algorithms, microtechnology, and multidisciplinary collaboration. These innovative technologies have already aided in disease detection for the commercial swine industry, surveillance for emerging viral diseases, and pathogen identification in ancient DNA, such as that from the time of the 14th century Black Death. In the future, LLMDA and LMAT will continue to drive pathogen-detection efforts for human medicine and bioterrorism.

—Lanie L. Rivera

Key words: algorithm, bacteria, big data, biodefense, bioterrorism, caching, Catalyst supercomputer, disease, dynamic random access memory (DRAM), fungi, gene, genome, k-mer, Lawrence Livermore Microbial Detection Array (LLMDA), Livermore Metagenomics Analysis Toolkit (LMAT), medicine, metagenomic, microbe, microarray, nonvolatile random-access memory (NVRAM), pathogen, sequencing, surveillance, virus, wound.

For further information contact Jonathan Allen (925) 422-0662 (allen99@llnl.gov).