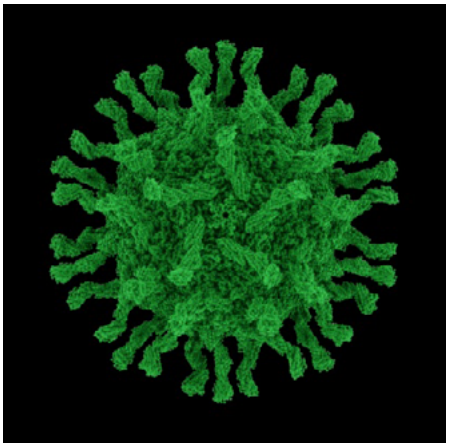
Centers for Disease Control and Prevention Automates Poliovirus Sequencing and Tracking



Poliovirus particle.

The Challenge

Support the Global Polio Eradication Initiative by tracking the transmission and evolution of poliovirus

The Solution

Use MathWorks tools to perform genetic sequencing, generate phylogenetic trees, and produce reports and maps used to help guide immunization programs

The Results

- Manual workflow automated and accelerated
- Cluster analysis time reduced by months
- Standalone sequencing tools developed

The existence of this story does not claim, infer, or imply CDC or United States Government endorsement, i.e. support of this MathWorks software tool over any other similar software, the MathWorks organization, or any other MathWorks product.

Eliminated throughout most of the world, polioviruses are still active in several countries, including Afghanistan, India, Nigeria, and Pakistan. The Centers for Disease Control and Prevention (CDC) supports the Polio Eradication Initiative of the World Health Organization (WHO) by providing epidemiologic and technical expertise to polio-endemic countries and partner agencies.

CDC's Polio Molecular Epidemiology Laboratory (PMEL) sequences poliovirus samples to determine their genetic signature and monitors the virus as it changes and spreads. The lab produces comprehensive reports that enable researchers to understand how the virus evolves during replication and to help health agencies mount more effective immunization campaigns.

MATLAB® and related toolboxes accelerate CDC's virus-tracking and reporting process. MathWorks tools enabled CDC to automate many steps in what used to be a labor-intensive workflow for managing and analyzing genetic sequencing data. As a result, CDC personnel spend less time on routine characterization and reporting tasks and more time on applied research.

The Challenge

The CDC lab processes patient data and sequences genetic samples from labs in West Africa. The polio results are compiled in a detailed monthly report for the WHO. Included in the report is a phylogenetic tree (dendrogram) that shows which viruses have been circulating in the area for the past three years and how they are related to each other.

In the past, producing this report and plotting poliovirus outbreaks geographically was a labor-intensive process that spanned multiple platforms and technologies, including Microsoft® Access™ databases and UNIX® based programs and scripts.

Assembling all the data for 3,000 sequences and then labeling, color-coding, and separating the viruses into clusters of genetic lineages took up to three days. The process was very complex, with a steep learning curve for cross-training others to do the task.

CDC needed to automate this workflow with tools that others in the lab could use and produce reports in a format that was easy to distribute and understand.

The Solution

MATLAB, Bioinformatics Toolbox™, and other toolboxes provided a platform for CDC to build tools that streamline the poliovirus tracking and reporting process.

To link patient data with individual strains, the researchers use Database Toolbox™ to read patient information, including the date and location of each genetic sample, into MATLAB, where they link it to sequencing information from a FASTA-formatted file imported using Bioinformatics Toolbox.

To analyze genetic data and identify clusters of genetically similar viruses, CDC researchers align genetic sequences and generate

MATLAB, Bioinformatics Toolbox, and MATLAB Compiler enabled CDC to streamline many manual steps within a single environment. A process that used to take three days can now be completed in hours, allowing labs to focus on the research that is so important to the polio immunization program.

neighbor-joining phylogenetic trees using Bioinformatics Toolbox and Statistics and Machine Learning Toolbox™. Working with MathWorks consultants, the team developed a MATLAB based cluster analysis tool that classifies viruses by serotype and genotype and then separates them into clusters of related viruses.

The team plots these clusters as color-coded dots on regional maps using Mapping Toolbox™. The cluster distribution maps enable health agencies to see where poliovirus is active and to detect patterns in the movement of the virus.

To simplify the overall workflow, CDC PMEL built standalone programs using MATLAB Compiler™. These programs feature an interface that makes it easy to select databases and files, annotate dendrograms with patient information, and generate monthly reports. More extensive documentation of annotated phylogenetic trees is produced using MATLAB Report Generator™.

In a related project, CDC researchers are studying how the poliovirus mutates and evolves. For example, they use MATLAB and Bioinformatics Toolbox to simulate mutations in the poliovirus genome over a period of 100 years. The results of this study will help health organizations understand how immunization programs may affect virus evolution.

The CDC PMEL is helping various specialized international polio research labs in Pakistan, India, and South Africa adopt the MATLAB based sequencing and analysis tools developed at CDC.

The Results

Manual workflow automated and accelerated. Producing the monthly polio report used to take three days. Using the tools that CDC built with MATLAB, Bioinformatics Toolbox, and MATLAB Compiler, any technician with minimal training can generate the report in about an hour.

Cluster analysis time reduced by months. In the past CDC researchers designated clusters by hand, writing on large posters and incorporating genetic difference data from spreadsheets. It was an immense effort spread out over three months. With the MATLAB based cluster analysis tool, all the data is in one place. The process is well-documented, and CDC researchers can complete it in one week of focused effort.

Standalone sequencing tools developed. The sequencing tools that the CDC polio group deployed using MATLAB Compiler will significantly improve the timeliness and communication of results within a region. Virologists in research labs that do not have MATLAB installed can use the tools to do their own mapping, label their phylogenetic trees, and pinpoint where viruses are appearing.

Industry

- Biotech and pharmaceutical

Application Areas

- Computational biology

Capabilities

- Data analysis
- Mathematical modeling
- Desktop and web deployment

Products Used

- MATLAB
- Bioinformatics Toolbox
- Mapping Toolbox
- MATLAB Compiler
- MATLAB Report Generator
- Statistics and Machine Learning Toolbox

Learn More About CDC

www.cdc.gov

MathWorks®
Accelerating the pace of engineering and science

mathworks.com