TOOLBOX

DEMOCRATIZING BIOINFORMATICS

Computational biologists are starting to develop platforms that open up the ability to analyse and interpret genetic-sequence data.

ILLUSTRATION BY THE PROJECT TWINS

BY JEFFREY M. PERKEL

For doctors trying to treat people who have symptoms that have no clear cause, gene-sequencing technologies might help in pointing them to a diagnosis. But the vast amount of data generated can make it hard to get to the answer quickly.

Until a couple of years ago, doctors at US Naval Medical Research Unit-6 (NAMRU-6) in Lima had to send their sequence data to the United States for analysis, a process that could take weeks — much too long to make pressing decisions about treatment. "If all you could do was get the data that you then have to ship to the US, it's almost useless," says Mariana Leguia, who heads the centre's genomics and pathogen-discovery unit.

But Leguia no longer has to wait for the analyses; she can get results in days or even hours — and she can do them in her own lab. Her unit makes use of EDGE (Empowering the Development of Genomics Expertise), a bioinformatics tool that hides common microbial-genomics tasks, such as sequence assembly and species identification, behind a slick interface that allows users to generate polished analyses. "We can have actionable information on site that allows us to make decisions very quickly about how to go forward," Leguia says.

EDGE isn't the first tool to simplify informatics with a point-and-click interface. Indeed, it lacks much of the flexibility and ▶

scope of more established alternatives such as Galaxy and Illumina's BaseSpace platform. But its simplicity is drawing in users who might otherwise shun bioinformatics. "People have used [EDGE] who would never have bothered learning command-line tools," says Clinton Paden, who uses EDGE in his work on virus pathogenesis at the US Centers for Disease Control and Prevention in Atlanta, Georgia. As such, it represents a case study in democratizing genome informatics — one that could help to accelerate uptake of the field by pure biologists.

INFORMATICS IN THE FIELD

Patrick Chain, who led the development of the software[1], at Los Alamos National Laboratory (LANL) in New Mexico, says that EDGE was created to try to square the rapidly growing availability of low-cost DNA sequencers with the relative paucity of know-how required to make sense of the data. It is designed for use in facilities that lack expertise in bioinformatics, says Joe Anderson, a computational biologist who honed the software for military applications at the Biological Defense Research Directorate (BDRD) at the Naval Medical Research Center in Frederick, Maryland.

It is also open-source, self-contained and provides end-to-end analyses for microbial genomics, from raw sequence reads to species identification and phylogeny in a single click. The system is also relatively cheap to run because the recommended hardware configuration (256 gigabytes of memory and 64 processors) can be bought for less than US$10,000, says Anderson. This means that most labs that can afford to run sequencing projects can afford the hardware. "That's not throw away money, but it's cheap enough," he says. It also helps that the set-up doesn't rely on an Internet connection and can be powered by a generator.

Users with reliable network connections can install the system to a cloud network. Nicholas Loman, a bioinformatician at the University of Birmingham, UK, points to CLIMB, the Cloud Infrastructure for Microbial Bioinformatics, which he helped to develop. CLIMB is a free service specifically dedicated to academics in the United Kingdom who are working on microbial genomics.

CLIMB was supported by £8.4 million (US$10.5 million) from the UK Medical Research Council and incorporates several informatics tools, including sequence databases and an analysis workbench known as the Genomics Virtual Laboratory. "I'm definitely thinking about having EDGE as a possible option on there as well," Loman says.

Overall, EDGE has been officially installed at 18 US Department of Defense and partner-nation labs, and on every continent except Antarctica, says Theron Hamilton, who is head of genomics and bioinformatics at the BDRD.

One of those is in Phnom Penh at the NAMRU-2 facility, which uses the system to track vector-borne diseases. "It's not traditionally the kind of place you would go to do bioinformatics," says Anderson. But EDGE is changing that. "One of the things I've realized is that, if you give [researchers] tools and get out of the way, they will amaze you," Anderson says.

The latest version of EDGE — version 1.5, released last October — includes 54 third-party tools. All components, including algorithms, databases, visualization tools and reference genomes, are housed on a server that drives six interlocking analysis modules: sequence clean-up; assembly and annotation; comparison to reference genomes; taxonomic identification; evolutionary analysis; and PCR primer design. Additional modules, including RNA analysis and pathogen detection, are slated for the upcoming EDGE 2.0, Chain says.

Last November, Chain and his colleagues demonstrated EDGE's capabilities in a study in which they used the platform to assemble, classify and map the evolutionary relationships in isolates of the bacteria *Bacillus anthracis* and *Yersinia pestis*; to untangle a mock human microbiome; and to analyse a series of human clinical samples, including cases of Ebola virus and *Escherichia coli* infection[1]. But the first published use of the system actually pre-dates that study by several months. Leguia's lab used EDGE to optimize methods for whole-genome sequencing of dengue virus — in a study published last June[2].

"People have used EDGE who would never have bothered learning command-line tools."

Users can explore those and other data sets using a free demo hosted on the LANL server. Researchers who wish to analyse their own sequences must install the software on their own systems. The code is freely downloadable from GitHub, and a Docker container and virtual machine image are available, but an information-technology expert will probably be required to handle the installation, says Chain. It is possible to tweak the source code to add other tools and workflows, but that's beyond the capabilities of many users, Chain acknowledges. A mechanism to simplify the process is in development, he says.

Paden, who has a background in computer science, says that the tool's simplicity makes computational biology accessible to researchers who might otherwise be intimidated by the usual tool for bioinformatics work — the computer's text-based command line.

But Titus Brown, a computational scientist at the University of California, Davis, warns that some of the benefits of EDGE are tempered by shortcomings that could limit the software's long-term use. He describes EDGE as an example of "opinionated software". "It gives you a small set of software to run that's been tuned to a specific set of examples," he says, "and it gives nice graphical summaries and outputs." But, he notes, it isn't clear how other researchers might help to improve the tool, nor what will happen should its funding dry up.

Chain says that the team made EDGE open-source partly because of concerns over future funding, which are also informing future development plans. "Sustainability is a question we have to think about," Chain says, "which is why we're going to try to allow third-party implementers to much more easily plug-and-play their projects, most likely using Docker."

A GALAXY OF TOOLS

EDGE is not the first bioinformatics system to offer a user-friendly interface. Galaxy, first published[3] in 2005, allows researchers to assemble informatics pipelines from a vast and flexible toolbox of free software offered through a web-based interface. Users can solve nearly any problem they can dream up by combining these tools in different ways.

But Galaxy can be intimidating to use. And, unlike the graphical representations generated by EDGE, such as phylogenetic trees or interactive 'Krona' plots of taxonomic data in hierarchical pie charts, Galaxy's output tends to take the form of processed data files, which the user then needs to take elsewhere to visualize.

"Galaxy is more like a kitchen, but there's no dining room," says Jeremy Leipzig, a software developer in the Department of Biomedical and Health Informatics at the Children's Hospital of Philadelphia, Pennsylvania. "The system is not really there for coming up with a way of delivering that output in an appealing way," he says. "With EDGE, they've actually thought about what the reports should look like."

Nathan Watson-Haigh, a bioinformatician at the University of Adelaide in Australia, says that EDGE could help to ease pressure on overworked bioinformaticians. But he cautions that it remains a complicated bioinformatics tool, and biologists who are inexperienced in computation would be wise to consult an expert before placing too much certainty in their results.

As with any tool, they need to understand what the algorithms are doing, and how different parameters affect their output, adds Kathleen Fisch, interim director of the Center for Computational Biology and Bioinformatics at the University of California, San Diego. "Just because you can run the tools doesn't mean that you should run the tools."

Still, as bioinformatics tools get ever easier, informatics could lose some of its aura of complexity. And for biologists, that could lead to wider adoption — and democratization. ■

1. Li, P.-E. *et al. Nucleic Acids Res.* **45,** 67–80 (2017).
2. Cruz, C. D. *et al. J. Virol. Methods* **235,** 158–167 (2016).
3. Giardine, B. *Genome Res.* **15,** 1451–1455 (2005).