Large-scale Scientific Computing in the Fight Against COVID-19

John West[1]

[1]Affiliation not available

January 4, 2021

Abstract

U.S. computing leaders, including the National Science Foundation, have partnered with universities, government agencies, and the private sector to accelerate research into responses to COVID-19 – providing an unprecedented collection of resources that include some of the fastest computers in the world. This current work expands on last month's Leadership Computing article by continuing to showcase the range of contributions that the national cyberinfrastructure is making to global efforts to stop the pandemic. This article touches on research efforts to learn how SARS-CoV-2 spreads among different populations, the biology and structure of the virus and its mechanisms of infection, and to develop effective vaccines for prevention and antiviral therapies for treatment. Even though we are still early in the process of developing an effective therapeutic response, the rapid mobilization of the national research cyberinfrastructure is a timely reminder of the strategic importance of robust, ongoing investments in large-scale scientific computing.

The COVID-19 High Performance Computing Consortium, first introduced in this space in the Nov/Dec 2020 issue of Computing in Science and Engineering (Hack et al., 2020), was established early in the effort to understand, treat, and ultimately prevent the sometimes devastating effects of COVID-19. The Consortium is led by the White House Office of Science and Technology Policy and brings together federal government, industry, and academic leaders who are making computational resources available to the COVID-19 research community. As of this writing, the Consortium provides 600 PFLOPS of computing capability via 6.8 million CPU cores and over 50,000 GPUs. (*COVID-19 HPC Consortium*, n.d.)

The National Science Foundation is a major partner in the Consortium, with many of its funded advanced computing organizations making significant contributions. These include organizations XSEDE, which manages request processing and allocations for consortium resources, the University of Texas at Austin's Texas Advanced Computing Center (TACC), the Pittsburgh Supercomputing Center (PSC), the San Diego Supercomputing Center (SDSC), and many others.

This article highlights several Consortium projects and extends the discussion in the previous issue by shifting focus to projects hosted outside the Department of Energy's Leadership Computing Facilities. We also broaden the discussion beyond the Consortium by bringing in additional supercomputing-enabled COVID-19 research to showcase the range of contributions that the national cyberinfrastructure is making to global efforts to stop the pandemic. This article touches on research efforts to learn how SARS-CoV-2 spreads among different populations, the biology and structure of the virus and its mechanisms of infection, and to develop effective vaccines for prevention and antiviral therapies for treatment.

Understanding the Biology of the Virus

SARS-CoV-2 was not known to science before the start of the global COVID-19 pandemic and was sequenced for the first time in January 2020, only about a month after its discovery. Since then over 100 organizations have contributed sequence data to the study of the virus that causes COVID-19, based on infrastructure developed during past efforts to sequence HIV, Ebola, Zika, influenza, Hepatitis C, and other viruses. The Galaxy project is one of the world's largest bioinformatics "gateways", supporting a community of more than 30,000 researchers, and is an important resource for analysis of the structure and stability of the SARS-CoV-2 genome. Galaxy users have access to a wide variety of computers, including TACC's Stampede2 and Jetstream supercomputers for large-scale computations, and the Bridges supercomputer supercomputer at the Pittsburgh Supercomputing Center for genome assembly.

Temple University's Sergei Pond has developed software called HyPhy for selection analysis in infectious diseases. Using Galaxy and HyPhy, researchers can perform analysis of SARS-CoV-2 genomic sequences. Because of the rapid availability of SARS-CoV-2 genome sequence data from infections around the world, researchers can use these to evaluate the degree to which the virus is – or is not – mutating over time. Thus far these analyses indicate that SARS-CoV-2 evolves slowly because of an enzyme that does error checking and correction during replication. This is important as less stable viruses make it difficult to develop effective vaccines.

Another science gateway being used for COVID-19 research is the I-TASSER gateway for automated protein structure and function prediction that is hosted at the San Diego Supercomputer Center. Yang Zhang, a professor of computational medicine and bioinformatics at the University of Michigan has been using the gateway to analyze sequences of the SARS-CoV-2 genome and compare them with coronaviruses in other species. Thus far these results suggest that pangolins, along with bats, may have played a role in the introduction of the virus to humans.

Discovering the Mechanisms of Infection

The first step in treating and potentially preventing COVID-19 infections is understanding how the virus infects its host cells. In general, scientists understand that, once inside an organism, the SARS-CoV-2 virus builds an extendable apparatus from core helical amino acids in its spike protein that latch on to a target host cell, leading to infection. However, if scientists can refine this general understanding into a complete picture of how the spike protein extends and then binds to its host cell, it may be possible to use the details of the process to find a way to disrupt the extension movement of the receptor-binding domain on the spike, preventing the virus from entering the cell and creating an infection in the first place.

Molecular dynamics simulations play an important role in understanding this behavior, but conventional methods are limited to simulation timesteps that are too large to develop a detailed understanding. Rommie Amaro's lab at the University of California, San Diego, is helping to accelerate the development of new treatments using the enhanced sampling weighted ensemble method on Frontera (Stanzione et al., 2020) to simulate the spike protein with the shorter timesteps needed to adequately resolve the relevant biological processes. As of this writing, TACC's Frontera is the 8th largest supercomputer in the world (*TOP500 List - June 2020 — TOP500*, n.d.), and its relevance to efforts to fight the pandemic indicate both the difficulty of the scientific challenges and the value of investments in leadership-class supercomputers.

Amaro's simulations have resulted in the discovery of important features of the virus, including the role that glycans play in camouflaging the virus from the immune system and revealing how the spike protein changes shape in a way that helps the virus bind with the ACE2 receptor on human cells.

Mahmoud Moradi from the University of Arkansas is using Frontera for simulations that study how the spike extension apparatus works, beginning with the observation that both SARS-CoV-2 and SARS-CoV (the cause of the 2002-2003 SARS epidemic) have spike proteins. Moradi's work relies on experimentally determined high-resolution 3-D structures of spike proteins in simulations of the features of both proteins, and investigates how the behavior between the two viruses differ. Thus far, the group has been able to observe significant differences in the dynamics of the binding mechanisms of the two viruses.

These kinds of numerical simulations are difficult and time-consuming, and reaffirm the unique value of leadership-class supercomputers like Frontera. In Baylor College of Medicine's Numan Oezguen's case, molecular dynamics simulations like these typically take 50 days of processing time to simulate one microsecond of viral action.

Disrupting the Ability of the Virus to Copy Itself

Once the virus binds to a host cell it hijacks that cell's replication machinery to make new copies of itself, furthering the infection in the target organism. If this process can be disrupted, it is possible that the duration and severity of infections can be reduced.

Scientists know from previous studies that the antiviral drug *remdesivir* interrupts the chemical processes the virus uses to copy itself by binding to enzymes responsible for the final assembly of copies. Andres Cisneros of the University of North Texas is using the Stampede2 and Frontera supercomputers at TACC to model the mechanisms that SARS-CoV-2 uses to copy itself to improve the effectiveness of antiviral treatments for COVID-19. His work investigates how remdesivir and other available drugs inhibit the proteins the coronavirus needs for replication. The key chemical reactions are simulated using a hybrid method called QM/MM (quantum mechanics/molecular mechanics) that speeds up the solution by using highly accurate calculations at the active site and using approximate molecular dynamics methods elsewhere.

Drug Discovery

Once SARS-CoV-2 has gained a foothold in an organism and is effectively copying itself to increase viral load, it is time for pharmacological intervention. Thomas Cheatham, a professor of medicinal chemistry and director of the Center for High Performance Computing at the University of Utah, is using Longhorn, an IBM/NVIDIA system at TACC to generate molecular models of compounds relevant for treatment of COVID-19. Once identified, the most promising candidates can then be tested in the lab in collaboration with medicinal chemists. Cheatham is using an approach he developed in 2015 to identify molecules for treatment of Ebola. The workflow selects promising amino acid chains and then conducts molecular dynamics simulations to optimize the structures. For COVID-19, the researchers are investigating the crystal structure of the main protease in the presence of peptide inhibitors. These simulations will then serve as the basis for laboratory experiments to identify the most effective candidates from the simulations.

Understanding the Spread of the Virus at Large, and Small, Scale

While the scientific community grapples with the urgent and difficult challenge of understanding the biology of the virus, its pathways of infection, and ways to effectively treat and prevent the COVID-19, the medical community deals with the devastating effects of the disease in patients on a day-to-day basis. Because the virus is new, effective approaches to manage and treat patients have been developed in real-time through trial and error. One area in which supercomputing is helping to fill knowledge gaps is in the understanding of transmission in hospitals and other indoor areas. Som Dutta from Utah State University leads a computational fluid dynamics project on Frontera to study how droplet clouds carrying the virus move and mix indoors. These simulations may help scientists develop procedures and guidelines to reduce droplet-based virus in a room, making it safer for health care professionals in contact with COVID-19 patients, and for other patients in the same facility. Dutta's simulations use multiphase large-eddy simulations to model the droplet cloud and to predict where particles will settle in an idealized hospital environment.

On a much larger scale, supercomputer models of virus transmission are an important tool for decision-making by local, state, and national leaders. UT Austin epidemiologist Lauren Ancel Meyers leads the UT

Austin COVID-19 Modeling Consortium, whose model is driven by anonymized mobile-phone data and case count and hospitalization data from Johns Hopkins University. They take an ensemble approach, combining results from two models to arrive at predictions. The first model fits a regression curve for daily death rates versus mobility data, and then makes extrapolations from that regression. This is not an epidemiological model as it does not make any attempt to describe the process of disease transmission. To account for this potential shortcoming, an "SEIR" epidemiological model is used as the second partner in the ensemble. The letters stand for the four categories of information used in the simulation: data on Susceptible (S), Exposed (E), Infected (I), Recovered (R), and Dead (D) patients. The key output of this model is each state's transition rate between S and E. In making a prediction for a state, both models are fit to the state's data, and the final prediction is a weighted combination of the two models. Ancel Meyers's model, and others like it, have proved invaluable tools for public health officials and policymakers struggling to contain the spread of the virus and to ensure the availability of critical care facilities to support patients experiencing the most devastating effects of the disease.

References

The U.S. High-Performance Computing Consortium in the Fight Against COVID-19. (2020). *Computing in Science & Engineering*, *22*(6), 75–80. https://doi.org/10.1109/mcse.2020.3019744

https://covid19-hpc-consortium.org. https://covid19-hpc-consortium.org

Frontera: The Evolution of Leadership Computing at the National Science Foundation. (2020, July). *Practice and Experience in Advanced Research Computing*. https://doi.org/10.1145/3311790.3396656

https://top500.org/lists/top500/list/2020/06/. https://top500.org/lists/top500/list/2020/06/