Cornell University
Center for Advanced Computing

CAC and John Bunge

CAC
we enable your success

Estimating Life’s Diversity on Land and At Sea

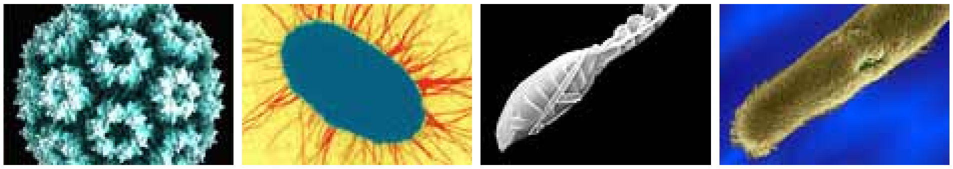
How many microbial species live in extreme marine environments? What is the estimated number of species in a given population?

Finding the Answer

CAC is helping Cornell’s John Bunge answer these and other biodiversity questions faster and more efficiently thanks to its high-performance computing systems and parallel computing expertise. Running his calculations in parallel rather than serially, Bunge has reduced his computing time by more than a factor of four.

Estimating Life’s Diversity

John Bunge is interested in using statistics to create accurate estimates of the number of species living in specific environments. His motivation is simple. With a better understanding of biodiversity, we will be better equipped to deal with environmental problems such as pollution and extinction rates, and to ask basic scientific questions about the diversity of life on earth.



Photos Courtesy of the American Society of Microbiology

Improved Research

Research Metrics

- Speed: Reduce compute time and research investigations by parallelizing code
- Efficient storage: Leverage SQL Server expertise and central storage capabilities of CAC for growing data sets with complex structures

Research Challenge

Because researchers can’t count every organism in a population, samples are taken. Population samples, however, often miss species. To account for missing species, statisticians plot the number of times each species is seen and construct a graphical curve. This curve is then extrapolated back in order to estimate how many species were observed zero times. That data point provides a basis for estimating the number of species missed in the sampling.

John Bunge is a leader in this statistical technique. He works with populations that contain hundreds, thousands, or even tens of thousands of species. “The idea is take a model of the count data and project downward into the invisible,” explained Bunge. “Then you add the zero count to the actual count and you get the total.”

In statistics, this process provides two results – the estimate of the number of species and a range of possible error. Getting the estimate within a reasonable error range is computationally intensive for a single data set. Bunge is now receiving multiple data sets, including 46 data sets of a microbial diversity survey that covers the entire planet.

Solution

CAC helped John Bunge reduce his research time by using parallel processing and the center’s HPC systems. “CAC deals with this scenario very well,” said Linda Woodward, a CAC consultant. “Each run is independent and can be done in parallel. While researchers can only run one data set at a time on a desktop computer, we can process multiple data sets simultaneously.”

Given the vast amount of data Bunge is collecting, he believes collaborating with CAC will enable him to develop SQL Server databases that will scale to meet his research needs and help him to effectively communicate and collaborate with the scientific community.

The Client

John Bunge, Chair of Social Statistics and Associate Professor in the Department of Statistical Science; Cornell University

- Species richness and biodiversity researcher
- Uses statistics to estimate number of species in a population
- Currently collaborating with biologists studying microbes
- Future research will focus on estimating numbers of expressed genes in DNA and paleontological diversity curves
- Bunge’s software is available to the scientific community

The Collaborative Relationship

“The Center for Advanced Computing took the pain out of doing the calculations. For me, it’s about research, about gathering data. Once it’s gathered, I want to analyze it and understand the insights that can be gleaned from it. The act of processing the data, getting from the collecting stage to the interpretation stage, can be time consuming and tedious. Shortening the process with CAC’s parallel computing resources lets me use my time much more efficiently.”

John Bunge
Chair of Social Statistics and Associate Professor in the Department of Statistical Science
Cornell University