Ohio Supercomputer Center

Data Consolidation

Carstens centralizes diffuse species data

■ NATURAL SCIENCES

Serving as a professor and vice chair of the Department of Evolution, Ecology, and Organismal Biology at The Ohio State University, Bryan Carstens has the privilege to share his passion for biology and genetics with students and fellow researchers. After years in the field, Carstens began to notice that imperative data sets containing information used for genetic research were often difficult to centrally locate. For example, if a researcher was studying a species of turtles, data about where the turtle samples were collected from may end up in a separate database than that of their DNA sequence.

"It makes sense that you would organize databases based on similarities. If it's a database for measurements of morphology, then that might be entirely different than a database for the genetic data," Carstens said. "But what it does is it makes it really challenging for other people to reuse that data."

Two challenges present themselves with this way of documenting data: researchers are hindered when performing large data analyses, and researchers and students lack easily accessible data about various species they may be studying or interested in throughout their careers.

To remedy this, Carstens has teamed up with Radford University and the Ohio Supercomputer Center (OSC) to build connections between different databases and package them in a way that allows for easy analysis and access using an analytical software such as the R programming language.

"Without OSC, I wouldn't be able to do the work or my lab wouldn't be able to do the work that we do. By the time we're done with this project, we'll have a database that has thousands of species' worth of data on it and then a set of different analyses that can be easily done without being a programmer," Carstens said. "We'll use Shiny R to make it very modular and to make it very intuitive and something you can do from a web browser. And so, by doing this, we're providing the kind of resource that hopefully will get lots of people excited about biology."

This project has been deemed the "Phylogatr," a contraction of the words phylogeographic data aggregator. Phylogeographic refers to the study of historical processes that led to geographic distribution amongst individuals, particularly in light of genetics. Additionally, this project is unique due to the fact that OSC employees are directly involved. Eric Franz, Shameema Oottikkal, Trey Dockendorf and Samir Mansour have all been involved, taking code written and tweaking it to extract the desired data.

The information that is being centralized consists of data that has been paid for and collected by the National Science Foundation over the last 40 to 50 years. This data is pivotal in both the research and academic realms. For students particularly, this development will lessen the time it takes to collect data and will allow them to research species that are of interest to them personally.

"Hopefully next summer or the following summer, we will release it to the public. The exciting thing is we're working with high school teachers from a local high school and we're hoping to tie it into the Ohio State curriculum for secondary education," Carstens said. "This allows lots of kids throughout the state to really use the supercomputer to do super cool analyses that they're designing on species that they're excited about." •

Project Lead: Bryan Carstens, Ph.D.,
The Ohio State University
Research Title: Phylogatr Project
Funding Source: The National Science Foundation
Website: carstenslab.osu.edu

Photo credit: The Ohio State University College of Arts and Sciences