

Race Car Code for Computational Biology

Performance optimization speeds algorithms for plant genetic studies



Variegated maize ears. Genetic difference in the maize genome lead to traits like coloration, growth rates, and hardiness. Scientists are trying to understand these genetic associations in order to create more productive crops. [Photo courtesy of Sam Fentress.]

Once upon a time, it was thought that genes were static. If you had this gene, you had brown eyes; if you had some other gene, you had a terrible disease. But scientists are increasingly realizing that most traits are determined by a complex network of genes working together to control the biochemical processes that determine the emergence of a trait.

“We’re now understanding that genes are literally like an online cybernetic control system that every minute of every day are turning each other on and off,” said Steve Welch, professor of agronomy[1] at Kansas State University. “It really is like a computer control system.”

Welch is part of a team using the pattern-detecting power of supercomputers to find important relationships among genes that may be responsible for traits in plants. The impetus for the project came from Pat Schnable at Iowa State University and Dan Nettleton, a statistician there with whom he collaborates. Their group was searching for pairs of genes involved in important traits. While the project is quite new, it is making important strides in terms of developing the methodologies for statistical association studies.

Few societal questions are as important as how to feed the world. Climate change, nutrient loss, and rising populations conspire to make this harder and harder. Scientists are racing to understand the genetic nature of draught resistance and insect hardiness so they can engineer tomorrow’s supercrops, capable of sustaining a growing human population.

“It’s getting cheaper every day to find out every letter in the DNA of all of the genes for many plants,” he said. “But it leaves you with terabytes of information you have to sort through and not just find the single genes that may be controlling traits, but to also look for the combinations. That’s the frontier right now: how do we start looking for combinations of genes?”

One of the main ways scientists “sort through” large datasets—whether genetic information or exoplanets—is by using high performance computers. In doing so, researchers transform their scientific problem into mathematical equations that are simulated via parallel computing.

The team developed an algorithm to find genetic associations, but early estimates suggested it would take 1,600 years using conventional hardware and software approaches to complete a simplified version of his problem. The project may not have progressed any further except for the performance optimization experts at the Texas Advanced Computing Center (TACC).

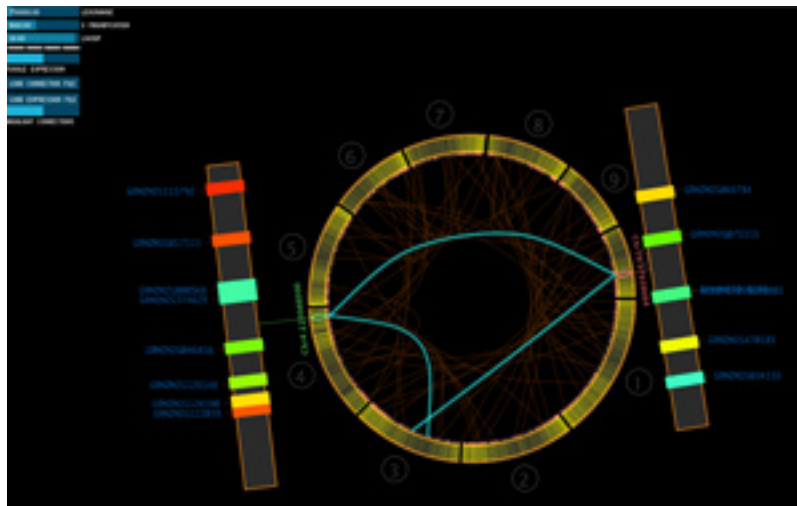
Schnable happened to run into Steve Goff, project director for iPlant, a large NSF program aimed at improving the tools for plant biology. Schnable told Goff about his problem, and Goff connected him with Welch and others involved with relating genes to their resulting traits. This included TACC staff, who not only run world-class supercomputers, but who are also experts at improving scientific computing codes.

Working with Lars Koesterke, a performance evaluation and optimization expert at TACC, the team simplified the mathematics of the problem, converted the code from Python to MPI (Message Passing Interface: a language for parallel computing), and got it to run on the Ranger supercomputer.

In doing so, Koesterke million times faster, reducing the time to to 4.5 hours, while at the number of iterations by improving the accuracy

Race cars and computer in common. Both are using intuition, trial knowledge. And both repeated loops.

"Most of the computing and you have to organize loop body is executed on the hardware," said that you exploit the optimally, and that's very the code correctly."



Potential epistatic interactions in the maize genome are revealed by interactive visualization of gene expression, genome structure, and the outcomes of the pairwise interaction application. Screen capture from iPlantInteractionBrowser (Matt Vaughn, TACC)

made the code run 3.2 according to Welch, solution from 1,600 years same time increasing the an order of magnitude, of the studies.

codes have many things created by engineers and error, and logistical go as fast as possible in

time is spent in loops, the loops so that the at the highest efficiency Koesterke. "That means hardware parallelism difficult, even if you write

As in a race car testing, small changes in the design can produce big changes in the outcome, or none at all. Treating software optimization like an engineering problem, Koesterke created several different kernels — the inner, most important loop of the code — where the logic was changed in each, to see how the order of different procedures impacted the speed.

“It was just trial and error,” he said. “If it runs faster, it’s better.”

Koesterke made the code run incredibly fast by eliminating unnecessary arithmetic and making aspects of the solution small enough to reside on the Level 2 cache (a shallow pocket of storage near the processor).

“If your code is three times faster, then you can solve a problem that is three times bigger,” Koesterke said. “But if your code is a million times faster, then you can do transformational science.”

The first round of analysis did not turn up any important associations. But, the innovations in the algorithm and code were significant enough to inspire a paper about the research that was accepted to the 2011 HiCOMB conference, an International Workshop on High Performance Computational Biology.

Association studies are hugely important, not only in plant biology, but also in studies of human health where they are expected to help uncover the genetic roots of most diseases. “Genes are genes,” Welch said. “Improved food, feed and fiber production are among the direct

