DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by Simbios, an NIH National Center for Biomedical Computing

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

Biomedical Computation Review
Simbios An NIH National Center for Biomedical Computing
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

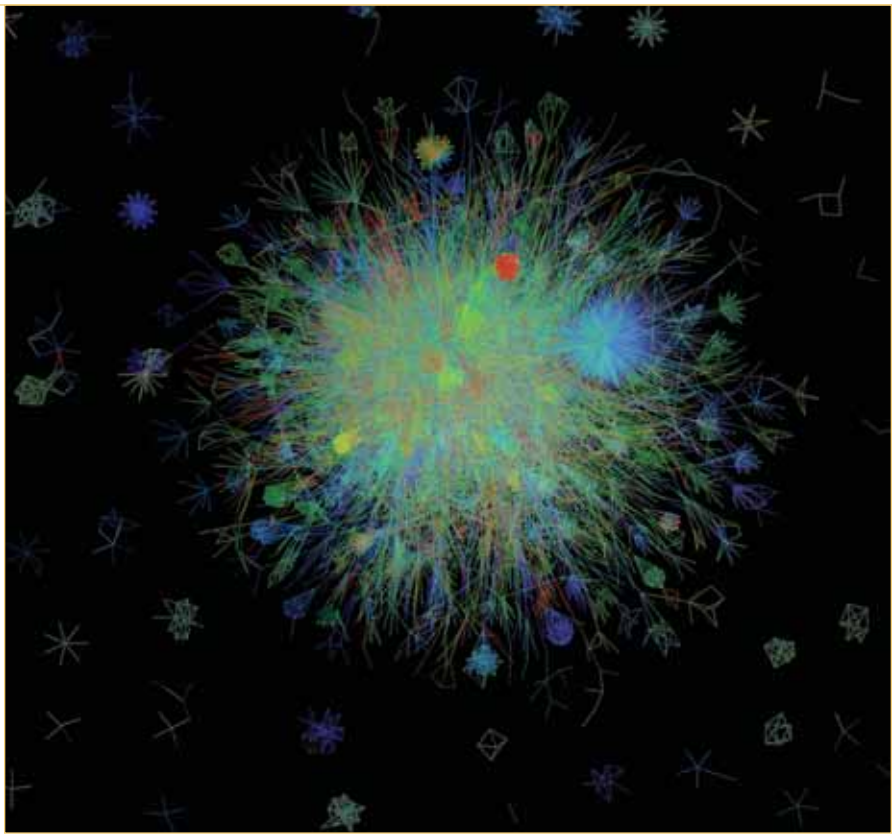
A Tipping Point for Function Prediction

There comes a tipping point in systems-biology studies of gene function where knowing some genes' functions can, using a computational approach, help hone in on the functions of other genes. That point has already been reached for yeast and *C. elegans* but is just now being reached for systems where functional information is more sparse—such as in plants and humans.

"There are still a lot of plant genes with unknown functions," says **Sue Rhee, PhD**, in the plant biology department at the Carnegie Institution for Science. "We need more sophisticated ways to characterize what these genes are doing."

So she and her colleagues, including **Edward Marcotte, PhD**, at the University of Texas, Austin, and **Insuk Lee, PhD**, at Yonsei University, South Korea, modified the *C. elegans* and yeast algorithm for use in systems with less complete data. This produced a

continued on page 29



This functional network of* Arabidopsis *genes shows the top 10% of the functional links identified by AraNet. Each line represents the connection between two genes and is colored to reflect the likelihood score for a relationship between the paired genes' functions: Red means a high score, blue is low. For example, the red area in the middle top of the figure represents the ribosomal complex, while the large blue cluster to the right represents the phosphatases, which have a weak relationship to one another although they share enough biological behavior to be linked. Image courtesy of Sue Rhee, Edward Marcotte and Insuk Lee.

```
float temp2;
{
  float multiplier = temp1;
  multiplier *= multiplier;
  multiplier *= multiplier;
  temp2 = multiplier;
  multiplier *= multipier;
  temp2 *= multiplier;
}
```

We are using only four multiplications to calculate a 12th power, which is much faster than the `pow()` function. Similarly, we can calculate the 6th power with three multiplications. But we can do even better by combining both of them into a single evaluation:

```
float temp2;
float temp3;
{
  float multiplier = temp1;
  multiplier *= multiplier;
  temp3 = multiplier;
  multiplier *= multiplier;
  temp2 = multiplier;
  temp3 *= multiplier;
  multiplier *= multipier;
  temp2 *= multiplier;
}
```

We are now calculating both powers at once with only five multiplications!

The final important optimization is to translate all expressions at once as a single unit. The above example shows only the expression for the energy, but in OpenMM we need to calculate the derivative of the energy as well. The two expressions share many subexpressions. For example, the derivative includes $(\sigma/\varepsilon)^{11}$ and $(\sigma/\varepsilon)^5$, so by translating both expressions together, we can compute four different powers at the same time.

In practice, we find these techniques work extraordinarily well for generating optimized OpenCL code to evaluate mathematical expressions. Our preliminary benchmarks with OpenMM show that the automatically generated GPU kernels are only a few percent slower than hand-tuned versions. At the same time, the user gains enormous flexibility to select the precise interactions they want in their simulations. □

Guest Editorial

cont'd from page 1

Every application that gets exchanged like this goes through CSR DRR.

While there is no central institute or office for biomedical computing and computational biology at NIH, there is a very vibrant and organic entity.

Now you're probably wondering about the outcomes of all these activities. In the last six years, the four broad-based BISTI announcements funded a total of 297 research grants in the amount of $355 million. In addition, the Continued Development and Maintenance of Software announcement funded 106 research grants in the amount of $160 million. In that same period, 5560 unique grant applications were reviewed in the informatics study sections (MABS, BDMA, BCHI, NT, GCAT, MSFD, BMRD, BMIT, MI and Continued Development and Maintenance special study section), and of these 1330 (24 percent) were funded.

For early stage investigators who want to add to these numbers by submitting successful grant applications, I offer the following advice:

- Team up with experienced mentors who can help you through the science and logistics of the NIH process.
- Talk to NIH program staff about your ideas. You can identify the appropriate contacts from the BISTI funding page/funding contacts link, http://www.bisti.nih.gov/funding/index.asp.
- Visit the BISTI Web site, which offers many useful resources, including a list of ongoing government programs, initiatives and public-private partnerships dealing with multiscale modeling, ontologies and data management, mathematical biology, systems biology, and numerous other biomedical informatics or computational biology efforts.
- Whether a new or seasoned NIH investigator, always focus your applications on the science because, after all, biomedical and health-related research is the NIH mission. □

Seeing Science

cont'd from page 30

rational approach to predicting gene function in *Arabidopsis thaliana*, a plant widely studied by plant geneticists. Dubbed AraNet, the work was published in the February 2009 issue of *Nature Biotechnology*. Marcotte and Lee are currently using the same approach to study gene function in humans.

"The idea is that we're making functional links between genes based on their behavior in a lot of different assays," Rhee says, including microarray analyses, protein-protein interactions and inferences from animal orthologs culminating in 24 different data sets.

The researchers started by analyzing pairs of genes with known function in order to set a baseline score for inferring related function. They then looked at about 27000 *Arabidopsis* genes—most of which are uncharacterized—to identify possible gene-gene associations among them. "By then asking 'what are the functions of the neighboring genes?' we can try to infer the functions of the uncharacterized genes," Rhee says. When her team experimentally tested the predictions for three uncharacterized genes, two out of the three had functions that were predicted by the network.

Rhee is interested in using inferences from AraNet to narrow down the candidate genes involved in complex traits. Although she'll be doing this work in plants, Rhee says the approach will be applicable to all organisms. She's also curious about uncharacterized genes that are connected only to other uncharacterized genes. "Perhaps we can use the network to characterize some undiscovered processes."

Ideally, Rhee says, researchers will combine AraNet's predicted functions with their own knowhow to try to design the best sorts of experiments to conduct. It's like rational drug design, she says: "You're using all the available information to be as systematic as possible in designing your experiments. This is a good application of systems biology." □