Gimme structure
by J. William Bell

University of Utah researchers use parallel genetic algorithms to predict crystal structures for a variety of organic substances.

NCSA has helped Julio Facelli do his work for years, and that gives him an uncommon perspective on the impact that the center and others like it have had.

“I’ve been following [the National Science Foundation-supported centers] since the beginning,” says Facelli, “and they’ve done a tremendous job of encouraging new approaches to science. In the beginning, users were the traditional suspects. But now more and more people realize that they need computational simulation to understand their experiments. In every field, the centers have had an impact on the way people do science.

“Simulation science is becoming part of the mainstream toolkit of the experimental scientist. The NSF centers drove that.”

Facelli, director of the University of Utah’s Center for High Performance Computing and a biomedical informatics professor, uses NCSA’s Mercury cluster to predict crystal structures for organic molecules that are frequently used in pharmaceuticals, fertilizers, and explosives.

“Fifteen years ago, the drug companies wouldn’t think about talking to a guy like me,” he says. “But computational molecular science—numerical analysis—can now show them the promise of a good solution for their formulation problems.”

Supercomputing is part of the reason for this shift, but so is the method that he and his team have developed for calculating crystal structures. This method was featured in a 2007 *Journal of Chemical Theory and Computation* publication.

**Survival of the fittest**

Crystal structure refers to the identical pattern of atoms that repeats over and over to form the macroscopic material. The details of this microscopic duplication have a big impact on the substance’s macroscopic properties, influencing features like solubility, reactivity, and color.

Modeling the crystal structure of a given substance, Facelli and his team begin with nothing more than the atoms in the molecule and the nature of their bonds. They’re looking for structures with the lowest energies, which typically mark the molecules’ standard crystal structures or something very close. The problem is that this straightforward data and straightforward goal create billions of possible solutions. Just imagine finding the needle of the lowest energy in that haystack of possible structures.

![Comparison of the experimental and MGAC predicted (pink) structure of L-alanine. The root mean square of the superposition is 0.30 Å, quantifying the close similarity between the experimental and the modeled results.](image-url)
To narrow the search to more likely candidate structures, the team uses a parallel genetic algorithm called MGAC that they have developed over the previous six or seven years drawing on systems at NCSA and on TeraGrid.

“An exhaustive search is not feasible, so we have to have a way to direct the search,” says Facelli. “[Genetic algorithms] are based on the principle of survival of the fittest. Trial solutions compete with one another in the population for survival and produce offspring for the next generation of tests. These algorithms offer excellent scaling properties, which make them good for large-scale parallel computing systems like those at NCSA and emerging computational grids like TeraGrid.”

For example, an initial calculation on NCSA’s Mercury may run 20 simulations on 20 different processors simultaneously, calculating possible crystal structures for a given set of atoms and their bonds. The energies for these structures are compared. The 10 with the lowest energies are kept, and the features of those 10 are mixed and matched to generate the structures for another 10 possible structures. Energies are calculated again, comparisons are made, best candidates are kept, and the cycle continues.

The “mating operation,” as the mixing and matching is called, stagnates quickly, producing very similar structures over the course of thousands of generations. To combat this lack of variety, the genetic algorithm also introduces arbitrary mutations into the process, occasionally taking one variable from one of the best candidates and including a random number for that variable in the next generation.

**From billions to hundreds**

Even with a genetic algorithm greatly reducing the number of candidate structures and the amount of time it takes to find them, the researchers are still frequently left with more than a million possibilities. To narrow it down further, they find and remove the many structures that are the same physically but have very slightly different profiles numerically, “essentially getting rid of the rounding errors,” Facelli says.

Next, they eliminate candidates that are the same structurally but that have revealed themselves in different orientations. This post-processing, which is done automatically at Utah’s Center for High Performance Computing, eventually gets the number of candidates down to a couple of hundred.
Every few years—four times since 1999—the Cambridge Crystallographic Data Centre hosts a "blind test" for the crystal structure prediction world. About a dozen participating research groups are supplied the atomic and bond information for a handful of molecules. They then use their different simulation methods to determine the crystal structure of the molecules, while an independent crystallographer works in secret to determine the real-world structures experimentally.

Sure there are some bragging rights associated with determining the correct structure or coming the closest, but the real value of the event is the experience and the collegial atmosphere, according to the University of Utah’s Julio Facelli. “It’s a very open scientific meeting [when they get together to compare results]. We discuss different methods, issue joint publications, and set goals [for the field] for the next three or four years. We discuss our results and the future.”

Facelli’s team used NCSA’s Mercury system to complete their entry in the 2007 blind test. Most of the methods used by teams at the test treat only fixed conformations of the structures, looking at flexible conformations by running each of the possible conformations of the structures individually. Facelli and his colleagues, however, hope that their more comprehensive method will identify the structures of flexible conformations without that extra step.

**Comparison of the experimental and MGAC predicted (pink) structure of N-(2-dimethyl-4,5-dinitrophenyl) acetamide. The root mean square of the superposition is 0.16 Å, quantifying the close similarity between the experimental and the modeled results.**

---

**The blind test**

“’It’s important to understand we’re not so much trying to predict the exact structure. There are dynamic factors influencing the crystal growth, which means that the experimental structure might sometimes be higher than the lowest energy,” Facelli says. Experts can compare those structures remaining and get it down to 10 or 20 that they want to test in the lab. “We take it from the billions of possibilities and say, ‘Here are the most probable.’”

*This research is supported by the National Science Foundation and the National Institutes of Health.*

**Access Online:** www.ncsa.uiuc.edu/News/Stories/gimmestructure

**More information:** www.chpc.utah.edu/~facelli/

**Team members:** Victor Bazterra  
Julio Facelli  
Marta Ferraro  
Matthew Thorley